

Science under attack

Researchers are increasingly upset with the Bush administration, not for its tactics but for its entire operational philosophy.

The highlight of the annual meeting of the American Association for the Advancement of Science (AAAS) last week was an impassioned session in which scientific leaders, including molecular biologist David Baltimore, made clear their views on the fraught relationship between science and the Bush administration.

The discussion was organized by the Union of Concerned Scientists in the wake of revelations about how the administration's political appointees have sought to control the messages communicated by scientists to the public, including attempts by the NASA press office to muzzle climate scientist James Hansen (see page 896).

And judging from the response at a packed and emotional hall in St Louis, a great many US scientists now believe that the Bush administration is prepared not only to ignore scientific facts in making policy decisions, but also to suppress findings that conflict with its own priorities.

For Baltimore — Nobel laureate, outgoing president of the California Institute of Technology, president-elect of the AAAS, and arguably the most eminent voice in all of American science — events have reached a tipping point. He suggested that the Bush administration's approach to science stems from its adherence to a particular philosophy of government, that of a 'unitary executive'. Instead of resignedly shrugging their shoulders whenever such a case of scientific manipulation arises, Baltimore argued, scientists need to recognize the potency of the threat that this governmental philosophy represents to the long-cherished independence of US science.

The unitary executive is an old idea, but not many Americans had heard of it until last month, when it cropped up during the Senate confirmation process of Supreme Court judge Samuel Alito. At the extreme, it holds that the executive branch can run the US federal government as it sees fit, especially in wartime. Given that a seminal achievement of the Constitution of the United States was to establish a balance of power between the executive branch, the Congress and the judiciary, this may sound absurd, but it seems to hold considerable sway within the Bush administration.

Baltimore warned that the doctrine opens the way for "an exertion

of executive hegemony over science". He called on researchers to "fight for a very different doctrine" under which "the executive's role is to defend intellectual freedom". In the light of the Bush administration's adherence to this philosophy, he added: "It is no accident that we are seeing such an extensive suppression of science." From someone of Baltimore's experience and reputation, these are strong words.

For science to flourish it needs settings that support freedom of enquiry, and the creation of such settings was a great achievement of the Enlightenment. Protecting them is vital, not just for science but for all of humanity.

Government agencies can never provide such settings in quite the same way universities can. But their scientists must still be allowed to express the results of their research as they see fit. They should also be free

to discuss how their research makes an argument for changes in policy, as Hansen sought to do with regard to climate change. In return, scientists have to acknowledge that the line between science and policy is a fine one, and endeavour to distinguish clearly between their scientific findings and their policy ideas.

In its five years in office, the Bush administration has sought to exert tighter control of the branches of government where scientists work. This applies not only to regulatory agencies, where politics are never far below the surface, but also to places such as the National Institutes of Health and NASA, where intramural researchers are used to the freedom of expression enjoyed by their university colleagues.

It is by no means the case that these proud federal agencies or their staff have fallen subject to the executive branch's decree. Most federal agencies have a deep stock of integrity, which even eight years of the Bush administration will not erode away. Yet Congress, in particular, should be doing much more to defend them from White House interference. And researchers should stand up and be counted with colleagues in the federal government in their hour of need. ■

"Scientists need to recognize the potency of the threat this philosophy represents to the long-cherished independence of US science."

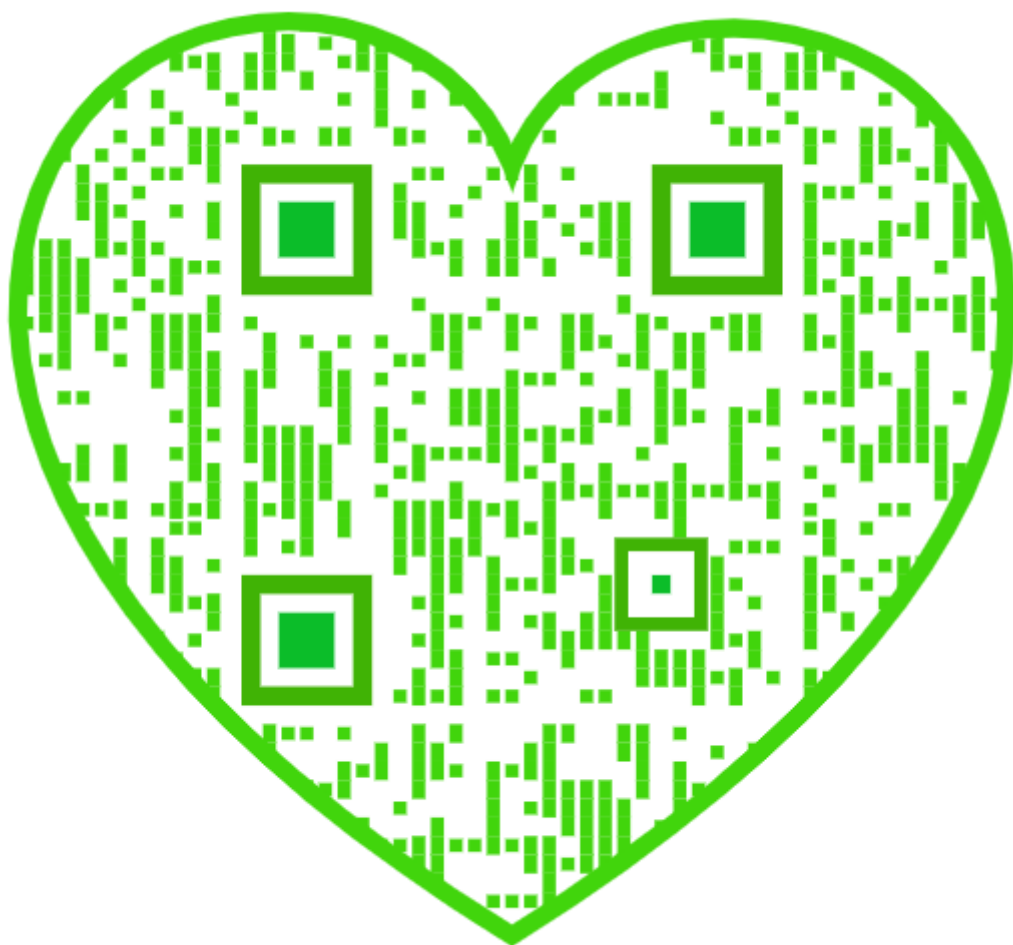
Not picture-perfect

Nature's new guidelines for digital images encourage openness about the way data are manipulated.

Researchers struggle to amass good data and present them in as clear a fashion as possible. But what do we mean by 'clear' when it comes to images? In days gone by, whether we liked it or not, data acquired at the bench were not much different from what was published. In a biomedical lab, for example, samples

that had been radio-labelled and separated on a gel were recorded on X-ray film. Composite figures were assembled, with lettering carefully placed around the mounted film. If a control was forgotten or a gel was uneven, the graduate student or postdoc was sent back into the lab to get it right 'for publication'. If a speck of dust on the film obscured data in the original photograph, another picture was taken. Slicing films to rearrange the order of samples, or to splice in a control group that was actually part of another gel, was not common because it took almost as much skill to do that as to rerun the experiment.

It is doubtful that scientists were more angelic then than now. It is



more likely that, when it came to image manipulation, they wouldn't because they couldn't. These constraints led to the accepted standards for publishing quality images: what you got is what you saw. It's not that researchers didn't aspire to perfection — to have obtained images worthy of the admiration of colleagues enhanced your prestige because it proclaimed technical mastery. But the numerous examples of slightly inadequate data continually reminded everyone just how difficult it was not only to perform the perfect experiment, but to acquire the perfect image.

Digital image acquisition and processing tools have removed the physical impediments to perfect images and laid bare the inadequacies of current training practices. Traditions for image handling were not passed down from one generation to the next because there weren't any traditions. Into this vacuum has crept 'beautification' — the digital manipulation of properly acquired data for the purpose of making a figure clearer, more perfect and more consistent with the best images yielded in such experiments. Removing dust from a digitized photo with the erasure tool, cropping bands from gels, and playing with fluorescence micrographs to enhance a particular effect are all attempts to show better results than were actually achieved in that run. In all these cases the data are legitimately acquired but then processed to yield an idealized image.

In *Nature's* view, beautification is a form of misrepresentation. Slightly dirty images reflect the real world. Accordingly — and after consulting with technical experts — the *Nature* family of journals

has developed a concise guide to appropriate image handling (www.nature.com/nature/authors/submissions/images), which will soon be incorporated into Guides for Authors.

In short, any digital technique that isn't applied to the entire image is suspect and needs to be explicated to the reader in the Methods or Supplementary Information. Authors should detail the instrument settings and the software manipulation that was performed on figures in an additional table in Supplementary Information. The fewer manipulations, the easier it will be for authors, referees and readers. Any changes made to misrepresent the data in the original image (such as boosting the contrast to eliminate backgrounds, making a collage of cells in a micrograph to show more cells in a visual field, or removing possibly informative bands in narrowly cropped gels) are of course strictly off limits.

We should conspire to end the fetish of the perfect image. Let's all get a little more 'real'. *Nature* is happy to work with others to aid the promulgation of image standards that we can all live with. The responsibilities of the institutes that train students, of the investigators who use their labour, and of the journals that publish the data can be better defined. Finding ways to regain our trust in scientific images is a goal on which we can all agree. ■

"In short, any digital technique that isn't applied to the entire image is suspect and needs to be explicated to the reader."

Diplomatic incident

Japan has fumbled its row with North Korea over tests to identify abductees.

Parties were held last week to celebrate the 64th birthday of North Korean leader Kim Jong-il — even in Japan, which has a troubled history with its near-neighbour. In Tokyo, the General Association of Korean Residents in Japan, which mainly comprises ancestors of Koreans who were forcibly taken to Japan during its 35-year occupation of the Korean Peninsula, held a party to mark the occasion. And those who attended were able to toast a recent diplomatic victory over Japan, in which the North Korean government was, incredibly, allowed to claim for itself the mantle of scientific objectivity.

The victory concerns the status of DNA tests conducted in Japan in 2004 on human remains that had been passed on from North Korea to Japan. North Korea said the remains were those of Megumi Yokota, a Japanese citizen that North Korea has admitted to kidnapping in 1977. Based on the DNA tests, however, Japan insists that the remains are those of someone else, and continues to demand an account of what really happened to Yokota.

But last January, in an interview with *Nature*, the scientist who carried out the DNA tests admitted that they were not conclusive (see *Nature* 433, 445; 2005). He has since been prevented from giving a full and open account of the matter.

Japan now needs to either produce evidence to back up its previous claims that the DNA tests were conclusive, or admit they weren't,

perhaps as a result of a lack of DNA in the incinerated remains.

But admitting such an error is something to which Japanese officials are inherently adverse. They may fear losing face — but such a loss would only be temporary, and coming clean would strengthen Japan's position in the longer term. There would still be no solid evidence to support North Korea's claim that Yokota was cremated. Japan could then continue to press North Korea for a plausible account of what really happened to Yokota and many other kidnapped Japanese nationals.

Instead, as things stand, the issue of the DNA tests on the bones has presented Japan with a thorny diplomatic problem. And the North Koreans are taking full advantage.

At bilateral talks two weeks ago, North Korea invited Japan to arrange a joint meeting of researchers to discuss the DNA analysis. Japan declined the offer, continuing to insist that its original interpretation was correct. This left the guests at Kim's birthday party gleefully dancing on their firm scientific ground: "We just want the truth to come out," they gloated. "We want to proceed scientifically."

Japanese officials need to learn from their mistake. In jostling with Kim's unpleasant regime, they need to be sure to retain the moral high ground. Public statements based on the wrongful interpretation of scientific data are liable to backfire. Refusing to acknowledge such problems at the first available opportunity tends to compound them, resulting not just in a loss of face, but of credibility. ■

"Japan now needs to either produce some evidence to back up its claims that the DNA tests were conclusive, or admit that they weren't."

RESEARCH HIGHLIGHTS

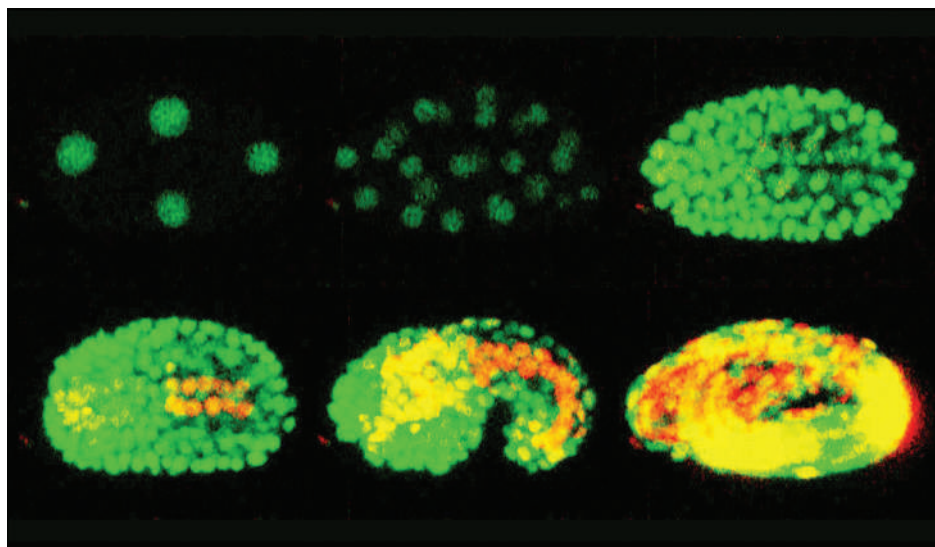
Follow that cell

Proc. Natl Acad. Sci. USA **103**, 2707–2712 (2006)

A cell-tracking system developed in living worm embryos could help to reveal the function of developmental genes.

Robert Waterston of the University of Washington in Seattle and his colleagues labelled the nuclei of *Caenorhabditis elegans* embryonic cells with green fluorescent protein and snapped 30 microscope images (examples pictured) every minute for eight hours as the embryo grew. The team built an algorithm that can figure out from these images which cell gave rise to which others. The algorithm takes just 25 minutes to run on a desktop computer.

The system can be used in combination with probes that monitor gene activity (appearing red in images).



A. WATERSTON, Z. BAO & J. MURRAY

CHEMISTRY

Fantastic plastic

Angew. Chem. Int. Edn doi:10.1002/anie.200504241 (2006)

Hydrogen cars need a safe and efficient way of storing the combustible gas. Could polymers be the answer?

Organic polymers are generally too floppy to form rigid, porous structures. But a recently reported class of organic polymer has molecules that contain bulky segments, opening up spaces in their packing. Neil McKeown of Cardiff University, UK, Peter Budd of the University of Manchester, UK, and their colleagues have now shown that such polymers can trap up to 1.7% of their own mass in hydrogen at -196°C .

That's still a long way from the US Department of Energy's goal to have a system that can store 6% hydrogen (by mass) by 2010. But organic polymers have great potential for structural fine-tuning, the team argues.

IMMUNOLOGY

One step to gene set

Nature Methods doi:10.1038/nmeth858 (2006)

A new kind of transgenic mouse should speed up the study of T cells, which help the immune system to tell 'self' bodies from non-self invaders.

Creating transgenic mice with different combinations of T-cell receptors can be a time-consuming endeavour, typically involving many rounds of cross-breeding.

Now Dario Vignali and his colleagues at St. Jude Children's Research Hospital in Memphis, Tennessee, have achieved this in just one step.

They used specialized retrovirus vectors to insert various T-cell receptor genes into mouse bone-marrow stem cells. These cells were then transplanted into T-cell-deficient mice, which started producing T cells with the desired receptors in five to eight weeks. Traditional methods would have taken six months.

GEOLOGY

Under ice

Geophys. Res. Lett. **33**, L02504 (2006)

Lake Vostok — a vast, ancient lake buried beneath the Antarctic ice sheet — has been found to have two similar neighbours (pictured below). Such lakes might harbour exotic forms of life.

The two lakes have areas of around 1,600 and 2,000 square kilometres. They seem, like Lake Vostok, to be controlled by local tectonics. This could mean that the lakes have been stable for hundreds of thousands of years.

To make the find, Robin Bell of Lamont-

Doherty Earth Observatory in Palisades, New York, and her team analysed satellite altimetry data, aircraft-borne gravity measurements and ground traverse data.

GENE THERAPY

Haemophilia on hold

Nature Med. doi:10.1038/nm1358 (2006)

Medical researchers have used gene therapy to provide a temporary cure for severe haemophilia B.

In a clinical trial of seven patients, Katherine High of the Children's Hospital of Philadelphia, Pennsylvania, and her team injected sufferers with a modified virus carrying the gene for blood coagulation factor IX (F.IX), which is mutated in haemophilia B patients. The gene was incorporated into liver cells and enough F.IX was produced to alleviate the uncontrolled bleeding that characterizes the condition.

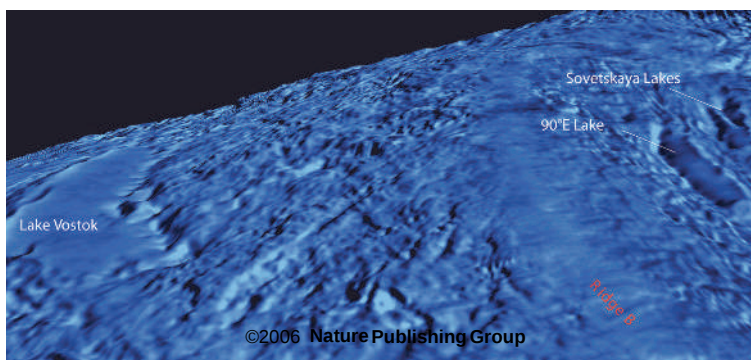
The effect lasted about eight weeks, after which the immune system is thought to have banished the modified liver cells. Using a different viral vector or administering immunosuppressing drugs may provide more permanent treatment, the researchers hope.

VIROLOGY

Fearful asymmetry

Cell **124**, 485–493 (2006)

A team of researchers in the United States has revealed how the unusual outer structure of the dengue virus



C. SHUMAN & V. SUCHDEO/NASA GODDARD CENTER; M. FAHNESTOCK/UNIV. NEW HAMPSHIRE

(pictured right) helps it to infect cells.

The hollow protein shells of viruses are made of protein units packed together like the panels on a soccer ball. In most viruses, each unit consists of a group of identical proteins arranged symmetrically. But the three proteins in the dengue virus unit turn out to be arranged asymmetrically.

Michael Rossmann of Purdue University in West Lafayette, Indiana, and his colleagues worked out how these proteins bind to a cell-surface protein known as DC-SIGN. Thanks to the asymmetry, DC-SIGN binds to sugar molecules on just two of the proteins, sticking the virus to the cell in a way that aids its invasion.

CANCER

Why the drugs don't work

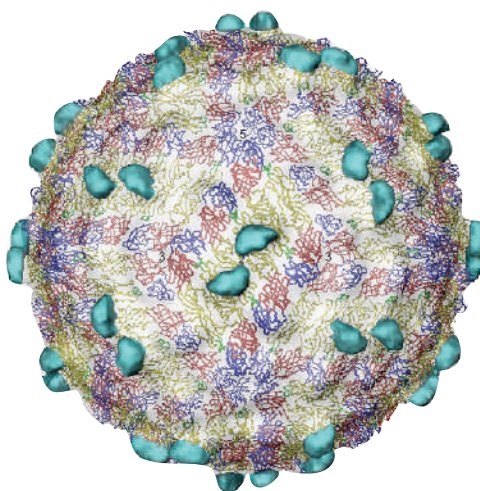
Cell **124**, 615–629 (2006)

Prostate cancer responds at first to drugs that block the action of androgens, such as testosterone, on which its growth depends. But resistance to the drugs soon develops.

Now Michael Rosenfeld and David Rose of the University of California, San Diego, and their colleagues report an unexpected mechanism that may explain why the drugs stop working.

They traced the effect to a chemical produced by macrophages — immune cells that infiltrate prostate tumours. The chemical produced, a cytokine, triggers a chain of reactions that ultimately reactivate the blocked androgen receptor.

The effect is mediated by a particular domain of the receptor. The same domain is found in all sex steroid receptors, and may have evolved for fine-tuning of hormone activity during sensitive events in mammalian reproduction.



QUANTUM PHYSICS

Clone rangers

Phys. Rev. Lett. **96**, 060504 (2006)

A quantum scheme dubbed 'telecloning' has been realized by researchers at the University of Tokyo in Japan, and the University of York in the United Kingdom.

Telecloning allows a quantum state to be simultaneously copied and reproduced in a number of remote locations. The protocol was implemented using three laser beams, entangled so that their quantum states were linked. First, a measurement was performed on one beam, the 'sender'. The result of that measurement was then used to guide manipulations of the two 'receiver' beams such that both adopted approximations of the quantum state specified by the sender. Quantum mechanics forbids the creation of perfect copies, but the experiments achieved copies almost as good as theory permits.

CHEMISTRY

At knife point

J. Am. Chem. Soc. doi:10.1021/ja057029t (2006)

Simulations suggest that antimicrobial peptides slice through the lipid bilayer of

microbial cell membranes like a knife through butter. Now chemists at the University of Michigan in Ann Arbor and the University of Massachusetts in Amherst, led by Gregory Tew and Zhan Chen, have demonstrated a way to watch the blade in action. The technique could help in screening new candidate antibiotics.

The researchers used sum-frequency generation vibrational spectroscopy to follow, in real time, the interaction of a typical antimicrobial peptide and a synthetic bilayer.

The team's observations of the lipid layer's destruction square with previous experiments, which have shown that the peptide inhibits bacteria when its concentration exceeds $0.8 \mu\text{g ml}^{-1}$.

DRUG DISCOVERY

Treating progeria

Science doi:10.1126/science.1124875 (2006)

A drug that interferes with a building block of the cell's nucleus could be used to treat Hutchinson–Gilford progeria syndrome (HGPS), a rare premature-ageing condition in which children rarely survive beyond their teens.

Researchers discovered in 2003 that HGPS is caused by a mutant form of lamin A, a key protein in the support scaffold of the nucleus. The mutant protein accumulates at the nuclear rim because a particular lipid section of the molecule behaves abnormally.

Loren Fong of the University of California, Los Angeles, and his colleagues show that drugs known as farnesyltransferase inhibitors, which block the addition of this lipid and are already in clinical trials as cancer drugs, reverse some of the symptoms in a mouse model of the disease.

JOURNAL CLUB

Michelle Peckham
University of Leeds, UK

A cell biologist turns detective to unravel the workings of one of the body's molecular machines.

I'm fascinated by the structure and function of myosins, a large, varied family of motor proteins. These molecules specialize in particular jobs — such as driving muscle contraction — and they move by walking along tracks

made of actin.

But understanding how myosins work has been like following a detective story, complete with false leads and unexpected twists.

One important characteristic of a myosin is whether or not it dimerizes; that is, pairs up with another myosin molecule. Dimers can use their two motor domains to walk along the actin, whereas monomers can take just a single step. Processive walking is useful for myosins carrying cargo long distances — such as the class that transports pigments in the skin.

Myosins can be zipped together into dimers by the α -helices in their tails wrapping around each other to form a structure known as a coiled coil. Unfortunately, the computer programs used to predict the formation of a coiled coil from sequence data are not infallible.

For example, for certain subsets of myosins, the programs predict that coiled coils should form, but in reality they do not. This is the case for myosin 6, which, contrary to predictions, has been observed (in some experiments) as a monomer.

So a recent paper showing that myosin 6 can dimerize after all comes as a surprise (H. Park *et al.* *Mol. Cell.* **21**, 331–336; 2006). The myosins pair up when densely packed on to actin, seemingly without a coiled-coil region, and do so more readily if the tip of the myosin's tail is missing. This hints that cargo binding to the tail could initiate the process.

But will the mechanism operate *in vivo*? And can other myosins dimerize in this way? Or is this just another false lead? I await the next chapter in the story.

NEWS

US scientists fight political meddling

ST LOUIS, MISSOURI

The rift between US scientists and the administration of President George W. Bush widened last weekend, as Nobel-prizewinning biologist David Baltimore used the annual meeting of the American Association for the Advancement of Science (AAAS) in St Louis to denounce government suppression of scientific findings.

Speaking last Saturday to a packed conference room, Baltimore — the president-elect of the AAAS — urged scientists to challenge perceived censorship of their research. Tensions between the Bush administration and researchers have been high for years, but Baltimore said he had recently grown convinced that the problem cannot be shrugged off as the usual battles between science and politics.

“It is no accident that we are seeing such extensive suppression of science,” he said. “It is part of a theory of government, and I believe it is a theory that we must vociferously oppose.” In particular, Baltimore condemned the “unitary executive” theory of government — the notion that a president can bypass Congressional and judicial oversight and run the country single-handedly (see page 891). Baltimore argued that this approach threatens to undermine the independence of science conducted under the auspices of the federal government.

Major US science agencies such as NASA, the National Science Foundation and the National Institutes of Health are all part of the



David Baltimore has called for opposition to the Bush administration's “suppression of science”.

executive branch of government, meaning that their employees answer ultimately to the president. In recent weeks, several researchers have gone public with charges that their government minders censored or otherwise manipulated their findings (see ‘Censored Science?’).

The latest round began last month, when James Hansen, a climate scientist at NASA's Goddard Institute for Space Studies in New

York, charged that NASA was trying to stop him doing media interviews that might cover policies on greenhouse-gas emissions. The 24-year-old NASA press officer who was the source of many of Hansen's complaints eventually resigned (see *Nature* 439, 643; 2006).

The accusations have left many government-funded researchers wondering about their role in public debate over science policy. Are they allowed to speak their mind based on the latest science? Or must they hold their tongue and respect their employer's wishes?

“There's no precise line that has been laid down,” says Daniel Greenberg, a guest scholar at the Brookings Institution, a non-partisan think-tank based in Washington DC. Instead, scientists must navigate the grey zone where science meets public policy.

Scientists employed by the government have different rights from those at universities or other private institutions, says Louis Clark, president of the Government Accountability Project, a Washington-based non-profit that is advising Hansen. Federal scientists, he says, can present their data publicly but must be clear that they are not representing their agency or government policy. “The cardinal rule,” Clark says, “is that you can speak for yourself, but you can't speak for the government unless you're authorized to do so.” Australia's largest funding agency, where scientists are also claiming to have been blocked from

Culture of fear reigns at Australian research lab

SYDNEY

Claims from Australian scientists that they have been gagged in discussions of climate change have revealed a culture of fear at the nation's leading research laboratory.

An investigative television programme aired by the Australian Broadcasting Corporation on 13 February claimed that three prominent climate scientists from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) were censored in discussions of climate change and energy research.

Against the backdrop of the Australian government's refusal to

back the Kyoto Protocol on climate change, there has been a public outcry over this and other media reports of political interference in the prominent research agency, which publishes high-impact studies on climate change. Political opponents have called for an independent inquiry into the allegations. On 21 February, the CSIRO announced a review of how it provides scientific input into policy development.

Graeme Pearman, a CSIRO veteran and former chief of the agency's atmospheric research division, says that in 2004 he was advised to withdraw his name from a report by the independent



In the streets: Sydneyites protest at Australia's refusal to ratify Kyoto.

Australian Climate Group that recommended targets for reducing greenhouse-gas emissions.

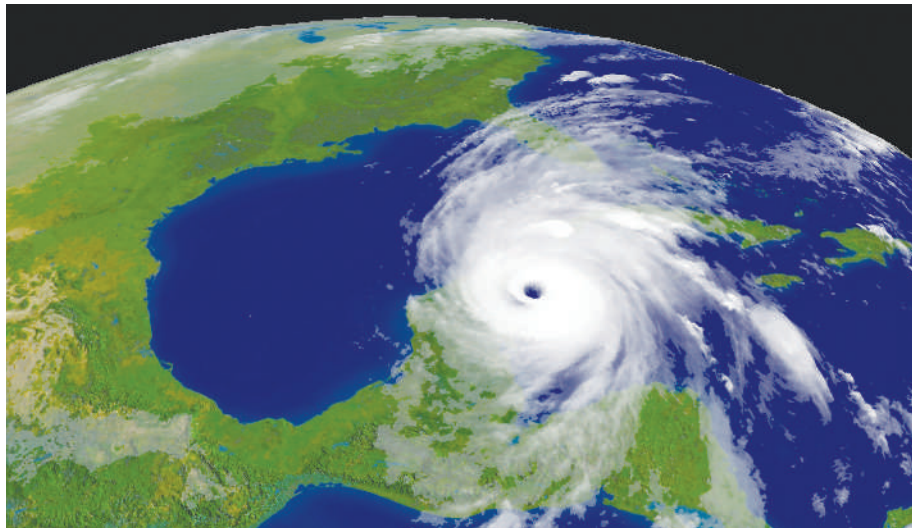
Steve Morton, chief of CSIRO Sustainable Ecosystems, and

Pearman's manager at the time, admits that he asked Pearman not to participate but denies that scientists are censored. “We encourage our scientific staff to



BIRD FLU HERE TO STAY?
Experts warn that H5N1 could become entrenched in Europe.
www.nature.com/news

NOAA



Hot topic: researchers have attacked official denials that hurricanes are linked to climate change.

talking about climate and energy research, makes a similar argument (see below).

Still, Hansen and others say that they must speak their minds for science to proceed. Hansen, for instance, is an expert on global temperature records, which are relevant to policies on greenhouse gases. "How can a democracy function," he asks, "if the public is not informed, honestly and fully informed?"

Such tales are not unique to the Bush administration. In 1993, physicist William Happer resigned unwillingly as head of the Department of Energy's Office of Energy Research after contradicting then-vice-president Al Gore. The vice-president had stated that ozone

depletion at the South Pole would lead to dangerous amounts of ultraviolet-B radiation; Happer pointed out that the oblique angle at which the sun hit the pole meant that even a thin ozone layer could absorb the rays.

But the current administration's policies on climate change, which most scientists agree are far out of step with the best available data, have exacerbated the situation.

Not all government employees have felt pressured. Rita Colwell, who headed the National Science Foundation under presidents Clinton and Bush, told the AAAS meeting that she had "not at any time come under political pressure from any quarter".

speak about their work. But there are certain areas where we urge them to tread warily and not to beat someone else's political drum."

Like US science agencies (see above), the CSIRO's official policy permits staff to discuss their scientific work, but advises against commenting on government policies. It says they should only express private opinions when it is clear that it is their personal view. But the line between science and policy is blurred, say CSIRO scientists, with the greatest tension being in environmental research into climate, water and energy.

"Scientists are encouraged to do policy-relevant research and yet not be policy prescriptive — there is a very fine line," says climate expert Barrie Pittock, another CSIRO

veteran who spoke out in the television programme. Pittock says he was told not to discuss mitigation in a 2003 report commissioned by the government's Australian Greenhouse Office. In particular, he says he was asked to tone down his discussion of how rising sea levels could displace people in southeast Asia, because of potential immigration implications for Australia.

Critics say the CSIRO's weakness in public debates on environmental issues in the face of conservative government policies is down to financial insecurity. Government funding has been eroded and there

is pressure on the agency to find external funding sources and industry partners. "There are a lot of fearful science managers second-guessing what they think government wants to hear," says the

third whistleblower, Barney Foran, who retired from the CSIRO last year. When the government appointed a biofuel taskforce in 2005, Foran says, he was warned not to say anything negative about ethanol fuel.

Other ex-CSIRO scientists deny being censored, but acknowledge that the mood at the agency has changed in recent years. "The CSIRO is very mute on key environmental issues at the

moment," says John Williams, who retired from the CSIRO in 2004 after heading the land and water division. And there is concern that younger staff feel constrained. "Anyone who has a baby and a mortgage would be crazy to speak out," says Foran.

But the culture of fear has been created by perception rather than reality, says Michael Borgas, president of the CSIRO's staff association. "The crux of the problem is job insecurity," he says, pointing out that annual staff turnover is 21% and that most new positions are short-term or casual appointments. "With that backdrop, people aren't going to be pushing issues or debates that bring the spotlight on them." ■
Carina Dennis

CENSORED SCIENCE?

Public-affairs officers are at the sharp end of charges that science is being suppressed or watered down by US government officials. Among the recent accusations:

- In October 2004, a NASA press officer was reportedly pressured by her boss to delay a news conference on ozone and air pollution until after the presidential election the following month.

- A University of Colorado sea-ice expert argues that NASA last autumn watered down a university press release to remove mention of accelerating sea-ice decline. NASA is re-evaluating its media policies.

- Scientists at the National Oceanic and Atmospheric Administration have protested at the agency's public position that denies any links between hurricanes and global warming. The administration recently updated its website to acknowledge that some researchers see a connection.

But speaking after Colwell, Susan Wood, a former scientist at the Food and Drug Administration, spoke of her reasons for resigning last August, after her boss repeatedly delayed a decision to make the Plan B contraceptive more widely available (see *Nature* 437, 179; 2005). The morale of scientists at her former agency was at its lowest point ever, Wood said. She got a standing ovation. ■

Colin Macilwain and Geoff Brumfiel

Y chromosome fuels dynastic dilemma

TOKYO

Politicians and legal experts debating whether to allow females to inherit the Japanese imperial line have consulted geneticists for advice. A government advisory committee considered whether the passage of an intact Y chromosome down a male line of descent might be a scientific argument for the male-only descent system. They concluded, however, that assigning rights to the throne on the basis of the Y chromosome would give a claim to many ordinary citizens.

The incident shows how desperate Japanese politicians are to resolve the problem of Japan's emperor, the latest in what may be the world's longest unbroken imperial line, having no male heir. When Emperor Akihito dies, the throne will pass to his two sons. But it's unclear what will happen after that. His grandchildren are all female, and at ages 39 and 42, the princes' wives are running out of time to produce a boy.

Prime Minister Junichiro Koizumi has thrown his weight behind an amendment to allow the oldest child of either sex to inherit the throne. This would allow the four-year-old daughter of the eldest prince to be next in line. Last year, Koizumi vowed to propose the amendment during the parliament's current session, which began on 20 January.

Nature has learned that in a series of cabinet meetings held ahead of Koizumi's pledge, politicians, historians and legal experts called on scientists for advice. Some experts, vigorously opposed to the idea of allowing females to pass on the line, used the Y chromosome to support their argument.

Government records show that committee member Hidetsugu Yagi, an expert in constitutional law at Takasaki City University of Economics, referred to the "stamp of the Y chromosome", and the need to preserve it.



Desperate for a boy: Emperor Akihito (second from left) is still without a grandson to inherit the throne.

Analogous pairs of all the other chromosomes swap genetic elements, which disrupts the integrity of the DNA sequences. But only 5% of the Y chromosome combines with the X chromosome; the rest is conserved. This means the Y chromosome has unmatched genetic integrity and forms an ideal marker for the male line (see "Tracing a Y chromosome through 100 generations").

"In a male-descent system, even a distant relative will have inherited the same Y chromosome," said Yagi. "Of course our ancestors didn't have this knowledge of genetics, but they would have known that blood can be inherited."

But the argument backfired, when Akinori Takamori, a historian of ancient Japan at Takushoku University in Tokyo, told the meeting that if the Y chromosome were so well conserved, it would also have been passed down a countless number of splinter lines over the past 1,000 years. "The distinction between the

average citizen and the imperial household would become uncertain," he says. Isao Tokoro, an expert in Japanese legal history at Kyoto Sangyo University, Kyoto City, agreed: "I don't think the 'stamp of the Y chromosome' is very meaningful here."

So the Y-chromosome argument did not make it into the committee's final report. After receiving the advice in November, Koizumi promised to push for a gender-neutral amendment in this season's parliament — a measure apparently supported by 80% of the population.

The situation was confused earlier this month, however, by the announcement that the second prince's wife was 6-weeks pregnant. With the possibility of a male heir within the imperial line, conservatives are pushing Koizumi to postpone the issue, something he looks set to do.

David Cyranoski

Tracing a Y chromosome through 100 generations

Those justifying a male-only system routinely refer to an unbroken male line of 125 generations since Japan's first emperor Jimmu. Although the first ten or so generations are probably mythical, and there may have been a break after the 25th emperor, Japanese historians say they are confident that the line has been continuous (if not always directly father-to-son) for 100 generations, or since about 500 AD.

If that's true, the imperial Y chromosome should be largely conserved, say geneticists. Accumulated point mutations are probably negligible, and although inversions and deletions might have built up, the sequence should still be recognizable, says geneticist Yutaka Nakahori of Tokushima University.

A genetics expert consulted by the government committee told *Nature* that digging up emperors to

analyse their DNA could show how far the Y chromosome stretches back, how much it has changed in that time, and whether the line has ever been broken. But he says he doesn't want to be named, much less involved in such a project, for fear of being "assassinated". And Mark Jobling of Leicester University, UK, who studies Y-chromosome evolution, says that most nuclear DNA from 1,500 years ago would have deteriorated: "You would be

very lucky to get good Y data from your ancient emperor."

An alternative to digging up graves would be to trace a tree of the imperial family by comparing the Y-chromosomes of distant living relatives to estimate when the oldest family member would have lived. Nakahori says he would love to do this, but realizes it would never be allowed by the Imperial Household Agency, which keeps a jealous guard on such matters. **D.C.**

Oil-rig staff get into marine biology

There used to be a drawback to Tony Kastropil's job. As a pilot for the robot submarines that monitor subsea drilling operations, he spent most of his time looking at murky underwater pictures of oil rigs. It was not an experience his friends and family had much interest in sharing.

But they have started paying more attention to the images he captures. Along with around 50 other pilots of remotely operated vehicles (ROVs), he now videos the often bizarre marine life he sees when studying rigs, and e-mails the results to grateful scientists. Thanks to Kastropil and others working off four continents, several new species and behaviours have been identified. And interest from rig operators is growing.

"ROV pilots see things I never get to see," says Ian Hudson, who coordinates the project from the National Oceanography Centre in Southampton, UK, pointing out that getting access to a ROV would normally cost tens of thousands of pounds per day. Hudson started the SERPENT project (www.serpentproject.com) in 2002, after a two-week trip he made to a BP rig produced enough data for three papers.

Realizing that the research potential of industry ROVs was largely untapped, Hudson got rig, ROV and oil operators, and the United Kingdom's Natural Environmental Research Council (NERC) to sign up to his project. The industry partners fly scientists out to the rigs, and together with NERC have created three PhD studentships for young scientists to take advantage of the opportunity. "There are probably 1,000

"The beauty as a pilot is that it gives you interesting stuff to relate to your family."

ROVs working for oil and gas companies at any one time," says Hudson. "The potential to explore is huge."

Some of the pilots have made significant finds since. Kastropil's team on the *Discoverer Deep Seas* in the Gulf of Mexico stumbled across the rarely seen dumbo-eared squid. Michael Vecchione, a marine biologist at the Smithsonian Institution in Washington DC, says videos of the creature sent to him through the SERPENT project helped him conclude that the animals belong to the Magnapinnidae family, adult

specimens of which have never been collected or properly described.

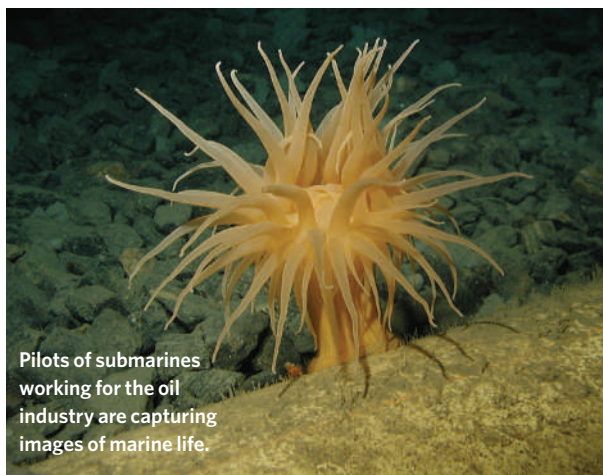
Hudson's personal favourite is a creature he dubs the 'murderous crab'. Galatheid squat lobsters (*Munida sarsi*) live at depths of 200–800 metres from Greenland to the Bay of Biscay. The lobsters were thought to scavenge fish remains from the sea floor. But in 2002, Hudson spotted one grabbing krill from the waters around it, before ripping off the animals' heads and devouring them.

The project is now expanding into the Arctic circle — Hudson expects to finalize a deal this week with Norwegian oil company Statoil. As well as giving scientists access to five rigs in and around the Barents and Norwegian seas, the firm plans to contribute nearly £500,000 (US\$867,000) over three years. Other industry partners, notably rig operator Transocean, supply a total of nearly £500,000 per year.

For many of the pilots, such as Kastropil, the project has awakened a keen interest in marine biology. The dumbo-eared squid was a "once in a lifetime" sighting, he says. "The beauty as a pilot is that it gives you interesting stuff to relate to family and friends. I've taken videos to my daughter's school. They were astonished."

Marine scientists have also praised the scheme. But some caution that it shouldn't be seen as a replacement for dedicated research vehicles. Chris Grech, deputy director of marine operations at the Monterey Bay Aquarium Research Institute in Moss Landing, California, says it's a great opportunity. "But the research will be severely limited by the dive locations, which are defined by the drilling needs," he points out. ■
Jim Giles

SERPENT PROJECT



Pilots of submarines working for the oil industry are capturing images of marine life.

Frosty US visa policy leaves Indian science cold

NEW DELHI

Scientific cooperation between India and the United States has been dented ahead of US President George Bush's official visit to New Delhi next month. Bush will find India's scientific community in a bitter mood following the United States' failure to give a visa to a leading Indian organic chemist on the suspicion that his work could be related to chemical warfare.

Bush was already preparing to deal with a nuclear establishment unwilling to separate its military and civilian atomic facilities the way Washington wants, a principle at the centre of last July's historic deal for nuclear cooperation between the two countries (see *Nature* **436**, 446–447; 2005). But the visa issue is creating a new wave of resentment.

Goverdhan Mehta, a former director of the Indian Institute of Science (IISc) in Bangalore, had been invited as a visiting professor to the University of Florida at Gainesville, but says he was asked to prove that he was not working on chemical weapons before a visa was issued.

Mehta does not seem to have been singled out — the US Embassy in New Delhi said in a statement that it is “obliged to obtain sufficient information necessary for evaluating each application”. But Indian scientists attempting to visit the United States on sabbatical or to attend conferences say that the eight-week ‘enquiry’ period is demeaning.

“This is too much,” says Mehta, who is a member of the Indian prime minister's science advisory panel and president of the Paris-based International Council of Science, which is committed to the free movement of scientists. “If the United States wants science cooperation with India it cannot humiliate our scientists the way I was treated by its consular officials.”

Many Indian researchers have now withdrawn from the process altogether. “As a matter of principle I have stopped going to the United States,” says biochemist Govindarajan Padmanaban, another former director of the IISc. “I am on the panel of the Indo-US



ICSU

Security risk? Goverdhan Mehta was asked to prove he wasn't working on chemical weapons.

vaccine action programme but have decided only to attend the panel meetings in India.”

C. N. R. Rao, science adviser to India's prime minister, says he agrees that visa applications need to be scrutinized, but doesn't see why it can't be done more tactfully. “I have decided never to attend an interview for a visa,” he says.

India's science secretary Valangiman Ramamurthi warns that the United States stands to lose out on collaborations. For instance, he claims that India's interest in the International Linear Collider project “has come down tenfold because our team could not attend a workshop in the United States due to visa problems”.

Samir Brahmachari, director of the Institute of Genomics and Integrative Biology in Delhi, says he was body-searched at an airport in New York last year, despite having a

diplomatic passport. “I have suggested that the IISc should cut off all US collaboration,” he says. “As for myself, I have decided not to go to the United States. My institute is looking to China, Japan and Taiwan for research collaborations.”

Officials at the US Embassy in New Delhi say it is not US policy to comment on individual visas. But in a statement released on 18 February, they said: “At the United States mission in India, certain cases involving high-technology issues require review before consular officers are authorized to issue a visa. This review does not in any way dilute the US government commitment to building a science and technology partnership with India.” ■

K. S. Jayaraman

“If the United States wants cooperation it cannot humiliate our scientists.”

ON THE RECORD

“Tell Larry he better appreciate my financial support.”

Harvard student Vikram Viswanathan joins an online betting pool on whether embattled university president Larry Summers will announce his resignation by 30 June.

“Native peoples were either living in harmony with nature or eating their way through a vast array of large-sized, attractive prey species.”

University of Utah archaeologist Jack Broughton questions whether early Native Americans were at one with nature.

Sources: *The Harvard Crimson*, *Reuters*

SCORECARD



Hawaiian holidays

The island of Maui may soon run out of sand. Mining sand for concrete is removing the only source for replenishing Maui's beaches.



Raves

Mixing the club drug ecstasy with loud music dramatically reduces brain function compared with popping a pill in peace and quiet, according to a study in rats.



The munchies

A drug that works by blocking a cannabinoid receptor in the brain may help to stop food cravings.

NUMBER CRUNCH

Is data secrecy hurting the next generation of life scientists? A recent survey of 1,077 doctoral students found that:

23% had asked for and been denied access to data associated with published research.

8% had denied other scientists access to data associated with their own published research.

51% reported that data secrecy had hindered their research.

Source: C. Vogeli et al. *Acad. Med.* **81**, 137–145 (2006)

Grizzlies, dodos and Gore put science on film

WASHINGTON DC

Al Gore, the former US vice-president who lost his fight for the presidency to George W. Bush in 2000, is starring in a new documentary that features his presentations on climate change. The film is just one of several science-related documentaries hitting the big screen this year.

The environment has long interested Gore, who published a book on the threat of climate change, *Earth in the Balance*, in 1992. More recently he has been touring the United States with a slide-show presentation on the dangers that his scientist supporters say is scientifically accurate and compelling. "He's very effective and right on target, and I've never seen better visuals," says Eric Chivian, director of the Center for Health and the Global Environment at Harvard Medical School in Boston, who helped Gore with sections on the effects of global warming on polar bears.

Michael Oppenheimer, a geoscientist and environmental-policy expert at Princeton University, agrees. "I have talked to thousands of politicians and very few of them have been willing to do the homework and get down to the details," he says. "When Al Gore gets revved up, he is very impressive."

But it wasn't until Gore's slide-show was made into a feature-length documentary, first screened at the Sundance film festival in Utah last month, that his efforts started making headlines around the world. *An Inconvenient Truth* was directed by Davis Guggenheim and produced by Participant Productions, a film company set up in 2004 with an explicitly activist agenda. The company's other films include *Syriana* and *Good Night and Good Luck*, which deal with the global oil industry and the freedom of the press, respectively.

Last week, Participant Productions secured a distribution deal with Paramount Pictures that should see *An Inconvenient Truth* get a US release in May. Supporters hope that the film will not only revive Gore's political career but will take his climate message to a wider audience than his original slide-show could ever have reached.

Thom Powers, a documentary producer who heads the documentary film programme at the Toronto International Film Festival, says that both science and political documentaries are becoming much more popular. "*March of the Penguins* was I think the second-highest grossing documentary of all time, second only



Al Gore (right) promotes his film on climate change while director Davis Guggenheim looks on.

to Michael Moore's *Fahrenheit 9/11*," he says. "Al Gore's documentary has a nice confluence of a couple of different interest areas: growing viability of political documentaries, and interest in the environment."

"Scientists have neglected the communication skills needed to engage with mass audiences."

The trend is growing. Moore, who helped get documentaries into movie theatres with his hit films *Bowling for Columbine*, about gun control in the United States, and *Fahrenheit 9/11*, about the Bush administration's actions in the wake of the terrorist attacks of 11 September 2001, is now working on a film about the pharmaceutical industry, called *Sicko*. And one of the year's most talked about films is Werner Herzog's *Grizzly Man*, a portrait of a man who lived among — and was eventually eaten by — Alaska's grizzly bears and claimed to be protecting them, although government experts say the population is doing well.

Doron Weber, who runs the Public Understanding of Science and Technology programme at the Alfred P. Sloan Foundation in New York, gave an award to Herzog's film. He says that boom times for documentaries are good for showing the subtleties of science. "A

film like *Grizzly Man* presents a character in the round and presents conflicting ideas," he says. "Film is such a good medium for this."

Another director taking advantage of the medium is Randy Olson. His film *Flock of Dodos*, about the clashes between evolutionary scientists and advocates of intelligent design, had its first screening in Kansas earlier this month. Olson comes down on the side of evolution, but sends a message that scientists have neglected the communication skills needed to engage with mass audiences, and failed to compete with the public-relations operations of advocates of intelligent design. "Science has backed into this lumbering, conservative corner, with no flexibility to deal with something like intelligent design when it comes up," says Olson.

Chivian says the same of the climate-change debate: "What really worries me is that scientists haven't done a great job of communicating these issues. We are trained to talk in technical language to each other." Many hope that the new wave of science documentaries will help to change that.

Emma Marris

Additional reporting by Kendall Powell.



CURLERS' GADGET SPURS STUDY OF ICE

Research team builds on electronic brooms to help cars stay on the road.

www.nature.com/news

Unrest returns to confront Harvard president

Harvard University's president Larry Summers is facing a fresh challenge to his leadership that some say could further undermine his ability to lead the institution.

Summers provoked a storm of controversy early last year after suggesting that intrinsic differences in ability between the sexes could explain why so few women have senior positions in science. The angry response mushroomed into a wider attack on what many describe as Summers' overbearing leadership style. It culminated in March in a vote of no confidence by the largest and most powerful of the university's faculties, the Faculty of Arts and Sciences (FAS; see *Nature* 434, 424; 2005).



Crisis of confidence: Larry Summers' leadership is again being questioned.

Summers weathered that storm, partly by apologizing repeatedly and promising to listen to his colleagues more. But anger among the faculty continued to simmer and has now bubbled over again. The latest uprising was triggered by the resignation in January of William Kirby from his post as dean of the FAS. Kirby disagreed with Summers on some issues.

"The faculty got very upset at the notion that the dean was being forced out," says Everett Mendelsohn, a Harvard historian of science and critic of Summers.

The brouhaha has triggered renewed attacks on Summers and calls for his resignation. The FAS is now set to vote on a second motion

of no confidence at a meeting on 28 February.

The ability to resolve the conflict may rest with the Harvard Corporation, the executive board that hired Summers and has the power to oust him. Last year, the corporation issued two statements in support of the university president — his bold style is said to be part of the reason they hired him — but it has so far remained silent on the current dispute. A further motion scheduled for vote at the 28 February faculty meeting calls on the corporation to intervene.

The search for Kirby's replacement may now be frozen until the situation with Summers is settled. The upheaval is also taking its toll on the university's ability to raise funds and is distracting from plans to reform the curriculum. "It's basically paralysing business above a certain level," says Richard Thomas, chair of the Classics department and a member of the faculty council.

Helen Pearson

M. HUMPHRIES/NEWS.COM

Ohio closes the door on intelligent design in schools

The Ohio state board of education has voted to remove a teaching plan from its curriculum that had been accused of favouring intelligent design.

Approved in March 2004, the tenth-grade lesson plan did not specifically mention intelligent design — the idea that an intelligent being shaped the course of evolution. But it did encourage a ‘critical analysis’ of evolutionary theory that included several standard arguments used by advocates of intelligent design.

In its 14 February vote, the board also struck out language requiring the teaching of a critical analysis of evolution from its state teaching standards. Many had criticized the language as opening the door to intelligent design or creationism.

The Ohio decision follows closely on from a federal judge’s ruling that the teaching of intelligent design is unconstitutional (see *Nature* 439, 6–7; 2006). Patricia Princehouse, an evolutionary biologist at Case Western Reserve University in Cleveland and director of Ohio Citizens for Science, says the ruling influenced the Ohio vote. “The decision made it very clear how this would be viewed in court,” she says.

Japanese company gives up contentious LED patent

A Japanese company has dropped its claim to a long-fought patent for blue light-emitting diode (LED) technology.

The patent was at the centre of a multi-million-dollar controversy between Shuji Nakamura, a materials scientist at the University of California, Santa Barbara, and his former employer, Nichia Corporation of Japan. In 2001, Nakamura sued the company, claiming that the patent was key to Nichia’s billion-dollar profits in blue LEDs and that, as its inventor, he did not receive the compensation guaranteed by Japanese law (see *Nature* 412, 844; 2001). In January 2004, a district court awarded Nakamura ¥20 billion (US\$169 million), based on an estimate of the patent’s value as ¥120 billion.



Shuji Nakamura: wanted compensation for patent.

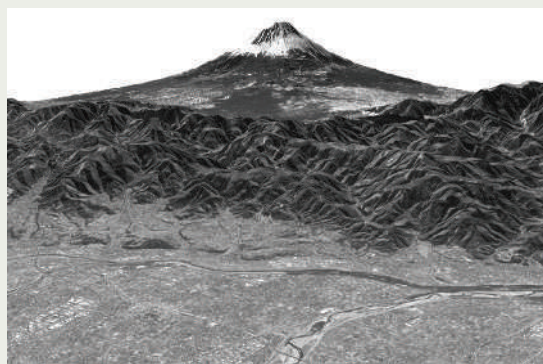
In an appeal to the Tokyo High Court, Nichia recalculated its view of the value of Nakamura’s contribution as only ¥10 million. In January 2005, Nakamura

Satellite tests its prowess by focusing on Mount Fuji

Japan’s sacred Mount Fuji towers here above the towns and rivers of the Kofu basin in one of the first images obtained by the Daichi Earth-observing satellite.

Launched on 24 January by the Japan Aerospace Exploration Agency (JAXA), Daichi can deliver high-resolution, three-dimensional images of Earth’s surface. This test picture was taken on 14 February using PRISM, an instrument with three independent systems that can simultaneously collect terrain and altitude data.

Other instruments on board allow Daichi to continue observing through any weather conditions, and to map vegetation at near-infrared wavelengths. JAXA says the satellite will help monitor natural disasters and identify natural resources in Asia.



JAXA

settled out of court for ¥840 million.

Earlier this month, Nichia announced that it would give up rights to the patent. “It wasn’t worth the cost of maintaining,” said a company spokesman.

Scripps Florida will be built away from wildlife refuge

After two years of legal and political wrangling, Florida officials have decided to relocate a planned research and biotechnology complex away from a sensitive wildlife area.

The Scripps Research Institute will now build its research complex near the campus of Florida Atlantic University in Jupiter.

In 2003, California-based Scripps, along with Florida’s Republican governor Jeb Bush, announced that the complex, housing some 550 scientists and technicians, would be built on a remote orange grove near a wildlife refuge (see *Nature* 431, 5; 2004). They heralded the project as a boost for Florida research, and Scripps put more than 100 researchers in temporary labs as planning moved forward.

But environmental groups sued, halting the development by arguing that there had been insufficient analysis of the project’s impact. After months of indecision, county officials chose the new location on 14 February. The planning process now begins afresh.

Greenland’s glaciers quicken their pace

Greenland’s glaciers are losing more than twice as much ice as a decade ago. Accelerated melting is causing some glaciers to flow out to sea at up to 14 kilometres a year. In total, glacial calving and melting sheds 220 cubic kilometres of Greenland’s ice

sheet each year, researchers reported last week in *Science* (E. Rignot & P. Kanagaratnam *Science* 311, 986–990; 2006).

Late last year, it was revealed that most of the continent’s ice cap is getting thicker, probably owing to increased snowfall. The middle portion of the island was growing at an average of 79 cubic kilometres per year as of 2003 (O. M. Johannessen *et al. Science* 310, 1013–1016; 2005). But the gains aren’t making up for the losses, the latest study suggests.

US academies to set up panel for stem-cell ethics

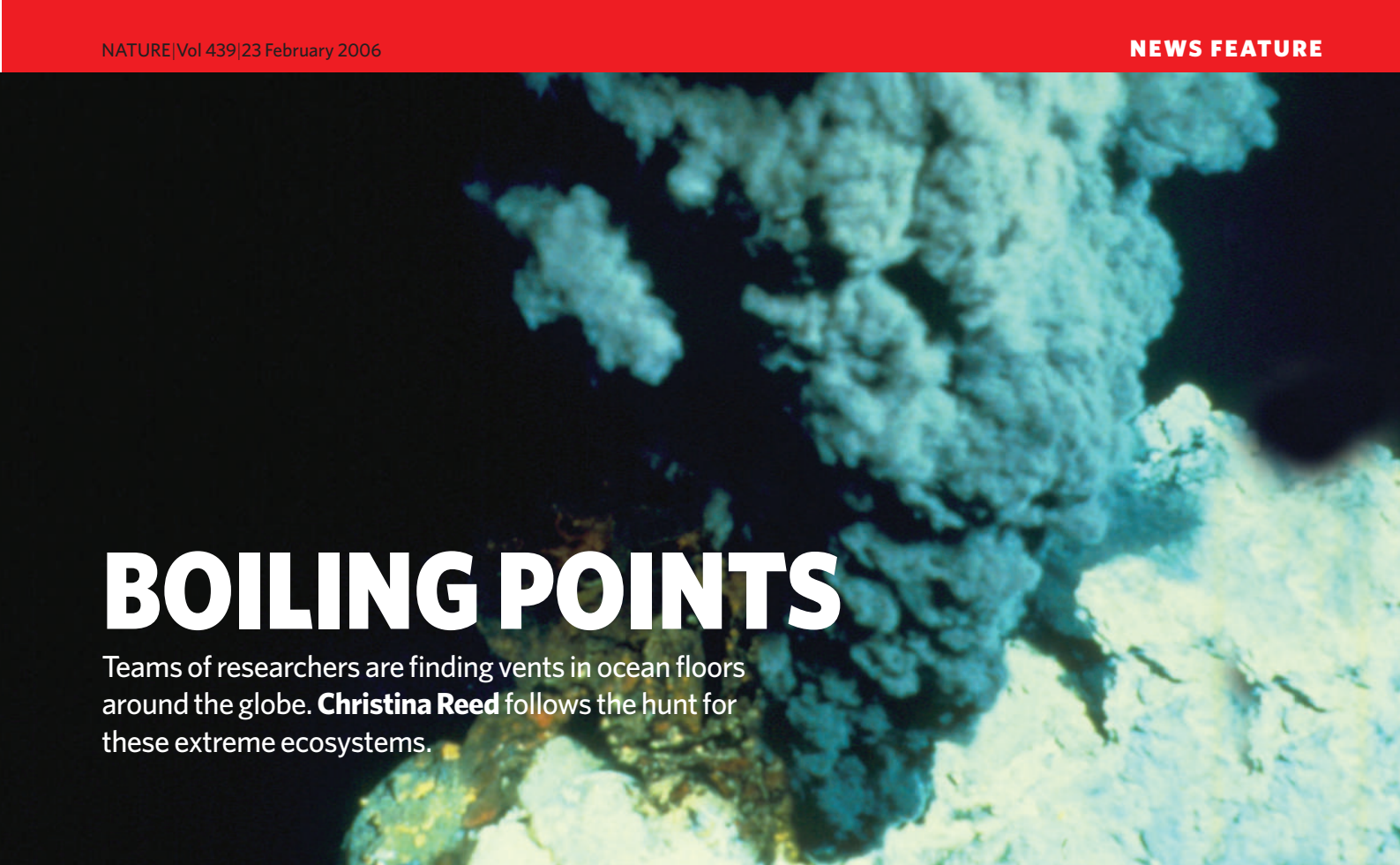
The US National Academies will create a committee to provide voluntary ethical guidelines for human embryonic stem-cell research. The decision is meant to bridge any regulatory gaps resulting from the restricted use of embryonic stem cells in federally funded research.

Currently, most embryonic stem-cell work in the United States is done by private companies, using regulations not specifically designed for this type of research. In the absence of federal guidelines, the academies last year issued its own, and the committee is intended to keep these periodically updated (see *Nature* 434, 1058; 2005).

The committee will be funded by private sources, including the Ellison Medical Foundation, the Greenwall Foundation and the Howard Hughes Medical Institute. The members of the committee have not yet been named.

Correction

The News Feature “Power struggle” (see *Nature* 438, 410–412; 2005) should have said that California governor Arnold Schwarzenegger has called for the state to slash its greenhouse-gas emissions to 20% of 1990 levels by 2050.



BOILING POINTS

Teams of researchers are finding vents in ocean floors around the globe. **Christina Reed** follows the hunt for these extreme ecosystems.

Before Rachel Haymon sailed to the Galapagos Islands last December, a fellow oceanographer made her a bet on whether she would find a particular kind of hydrothermal vent. The Galapagos region is where scientists discovered these vents, underwater fissures where superheated water bubbles out of the sea floor. But one of the most dramatic vent phenomena — black smokers, chimneys that spew a flood of dark mineral particles — had never been seen in the area.

According to some, this was because the geological setting of the Galapagos should not permit the formation of black smokers. But in retrospect, Haymon says, she should have upped the ante on her bet. On 14 December her team discovered the first Galapagos black smoker.

And that isn't the only new find. In the past couple of years, marine scientists have discovered numerous vents. What was once a handful of isolated vents has expanded into a dizzying diversity of oceanic wonders. Researchers don't yet know what to make of all the finds, but for now they are sitting back and enjoying the feast.

The recent finds include the most northerly active high-temperature chimneys in cold Arctic waters; chimneys in the South Atlantic; and the biggest plume ever recorded of hydrothermal chemicals released from an underwater eruption. With each new discovery, marine scientists are realizing that the unusual is to be expected. Variations in geology, chemistry, physics and biology

make every vent field unique.

Take, for instance, the warm water flowing from hydrothermal vents in the chilly waters of the Arctic Ocean. Here, 500 metres below the surface, iron-oxidizing bacteria, including the species *Gallionella ferruginea*, cover the sea floor for kilometres. Within this region, known as Gallionella Garden, short mineralized chimneys sprout up, each around 15 centimetres high and often topped with delicate sea lilies (*Heliometra glacialis*) that make them look like one of Dr Seuss's cartoons.

In July 2005, a team led by Rolf Pedersen of the University of Bergen, Norway, visited this

"When you consider the size of the ocean and how hidden everything is, we've only begun to really know what's there." — Rachel Haymon

site. It lies along the Mohns Ridge, where tectonic plates are pulling apart, or 'spreading' at a rate of about 16 millimetres per year; speedier ridges can spread by up to 200 millimetres per year. The researchers sent down a remotely operated vehicle called *Bathysaurus* to look at the region. Using temperature sensors, the group discovered that the chimneys bathed the sea lilies with a flow of fluid typically at 0.5 °C. That's warm for the Arctic Ocean; the background temperature at this site measured – 0.3 °C.

The team then sent *Bathysaurus* to the base of an escarpment, formed when an earthquake pushed up the sea floor 100 metres. Following

the base of the cliff wall, the submersible found a series of hydrothermal mounds, each topped with several smoker vents. The fluid pouring out of the smoker chimneys here was far hotter. Because it was bubbling, the team estimated it must be at least 263 °C. And sure enough, when *Bathysaurus* extended its 260 °C-maximum sensor into the flow, the probe was knocked out of commission.

Smoking chimneys

While hovering over the site, the research team noticed something unusual just below their ship. "The echo sounder recorded something coming up," says Pedersen. He wondered whether the echo sounder had detected fish that had followed a drift of nutrients wafting up from vents on the sea floor. Using a similar signal farther down the ridge, the team discovered another vent field within hours, just 5 kilometres away from Gallionella Garden.

After years of laboriously tracing chemical plumes to locate vent sites, the scientists were thrilled with the ease of their discovery. The echo sounder not only helped them locate the vents, but also imaged the plumes and what appeared to be an associated school of fish. "This had never been done before," says Pedersen. Next summer, the team will return to the vents to sample fish and other animals, microbes and fluids, and then use the new technique to search for vent sites in deeper waters farther north.

As some researchers push north, others are going south — finding hydrothermal vents on the southern part of the Mid-Atlantic Ridge.



Hot spots: minerals and heated water spewed out of hydrothermal vents may support marine life that wouldn't otherwise survive.

This is the underwater mountain range that runs like a backbone down the middle of the Atlantic Ocean. As many as two dozen hydrothermal-vent fields have been found in the northern part of this mountain chain. But few expeditions had searched for vents in the south. So it was with an eager crew that the RSS *Charles Darwin* embarked on a month-long expedition in February 2005, to map the sea floor along a 250-kilometre stretch of the ridge.

Near 5° S, the team found two potential vent sites a kilometre apart, and mapped each with high-resolution sonar using an underwater robot named ABE (for autonomous benthic explorer vehicle)¹. On one of ABE's trips, its mission was to photograph the field of potential black smokers with a digital camera. "When we brought ABE back on deck, it was burnt through," says the expedition's leader, Chris German of the National Oceanography Centre in Southampton, UK. ABE's images showed that "the cooking place", as German calls it, contained fresh basaltic lava flows and what looked like shrimps, mussels and crabs.

In April, the German research vessel *Meteor* visited three sites in the area. At one site, called Turtle Pits, a team led by Karsten Haase of the University of Kiel, Germany, sampled boiling fluids at temperatures close to 400° C.

"This temperature is the highest measured so far along the entire Mid-Atlantic Ridge," says team member Andrea Koschinsky, a geochemist at the International University Bremen in Germany. The high temperature indicates that fluids are being heated close to a magma source, most likely from a recent

upwelling or volcanic eruption. "The effect of recent volcanic activity on submarine hydrothermal systems has so far been documented only along fast- and intermediate-spreading centres, such as the East Pacific Rise, but not from slow-spreading ridges where volcanic eruptions are rare," she says.

The German team discovered another vent field at 9° 33' S, the most southerly vent field known so far on the ridge. Because very young and small mussels dominate the ecosystem, the group dubbed the field 'Liliput'. In April, Koschinsky will lead the next *Meteor* cruise to the region, with a freshly mended ABE ready to track down any more hot vents.

Oceanographers increasingly realize that when it comes to vents, they don't know what is out there. So they have taken to using plume sensors whenever possible — even when mapping a suspected plate boundary. Such was the approach of marine geologist Bramley Murton, when on board the *Charles Darwin* in the Indian Ocean in 2003.

Plume clues

Somewhat optimistically, Murton attached a series of sensors called miniature autonomous plume recorders (MAPRs) to the tow-line of a dredge he was using to conduct a geochemical survey of the slow-spreading Carlsberg Ridge. Geologists only thought the ridge existed because Russian magnetic mapping of the sea floor in 1970 had predicted it should be there — not because anyone had actually seen it. Murton's expedition was the first to map the ridge with multibeam swath bathymetry, a kind of sonar that produces beautiful high-resolution images

of the seafloor topography. The MAPRs were attached to the dredge, just in case, to pick up any hints of a plume from a submarine eruption.

Murton's team managed to map the ridge, and excitedly pulled up glassy basalt rocks from it, but the biggest surprise was the MAPR results. The sensors had detected, in the chemistry of the water, evidence for a recent underwater explosion. Further lab tests, the results of which were announced at the American Geophysical Union meeting in San Francisco last December, showed that they hadn't come across just any hydrothermal plume. It was an 'event plume', meaning it came from a single underwater eruption.

Most plumes that oceanographers use to track down vents come from a single chimney or vent field. If the chemical signal from a vent field is like the smoke from a candle, that from an event plume is like the smoke from a wildfire. Even more surprising, this one turned out to be the most energetic and biggest event plume ever seen — 7 to 20 times larger than those previously recorded. It had risen 1,400 metres above the sea floor, measured one kilometre wide, and was drifting for 70 kilometres along the ridge. Twenty million cubic metres of lava would be needed to create the heat to drive such a plume, says Murton. Clearly it came from a major volcanic eruption.

Because the Carlsberg Ridge is one of the slowest-spreading, and so supposedly less active oceanic ridges, many had thought it unlikely to be the location of a major volcanic eruption. At ridges such as this, heat is thought to be released more slowly from the underlying magma.

Such oceanographic revelations show the

importance of exploring new areas. But the search doesn't always have to involve month-long, deep-sea expeditions that typically cost millions of dollars. Scientists are increasingly seeing hydrothermal vents through a dive mask rather than through a submersible's viewport.

Because the sea floor is on average about 4 kilometres from the surface, hydrothermal vents are frequently referred to as 'shallow' if they are anything less than 1 kilometre below the surface. But truly shallow hydrothermal vents, the kind a researcher can scuba dive to, are very different from their deep-water counterparts, as recent studies have shown.

Deadly cocktail

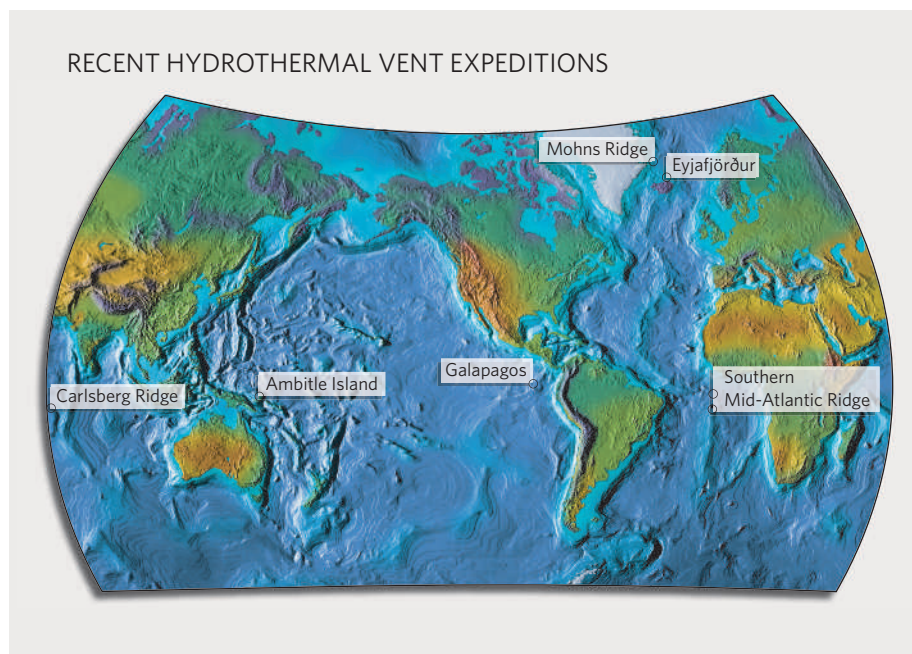
In an Icelandic northern fjord called Eyjafjörður, researchers discovered in 2004 a rich shallow-water vent ecosystem. "The area contains not a single vent, but a series of chimneys and fissures stretching for about 500 metres," says geologist Bjarni Gautason, who works for the company Iceland GeoSurvey. The tops of some of the chimneys are only a 14-metre dive from the surface. All the chimneys spout fresh water, contaminated with less than 1% sea water, and at temperatures reaching 77 °C. The chimneys "form their own landscape providing a unique habitat for the animals and plants in the area," Gautason says.

"We know about some other shallow-water hydrothermal-vent sites in the world, but none that makes chimneys like the ones here," says biologist Hreiðar Valtýsson of the University of Akureyri in Iceland, and a member of Gautason's team. "So we think ours are quite special."

But shallow vents aren't always the kind you would want to visit. Near Ambitle Island, Papua New Guinea, some vents spew out a toxic cocktail of arsenic with a splash of sea water. As much as 1.5 kilograms of arsenic flows daily from these vents into the surrounding coral reefs².

For up to 200 metres around the vents, microbes coat the sediment and corals with a red and green biofilm. These organisms seem to be able to survive the contamination, and so may give researchers a chance to study arsenic in a marine environment, and see how microbes cycle it through the ecosystem. In the past few years, studies have suggested that arsenic from silt is getting into the water supply in Bangladesh and is potentially poisoning millions of people. The vents near Ambitle Island could provide clues on how to address such a problem.

In Bangladesh, arsenic is being reduced, by microbial activity, from As(V) in the sediment to the more soluble As(III), which seeps into the groundwater and contaminates the drinking water. Jan Amend, a microbial geochemist at



Washington University in St Louis, hopes that his team's work at the Ambitle vents will help researchers better understand how microbes might catalyse the reverse chemical reaction, oxidizing As(III) back into the less soluble As(V). "This part of the arsenic cycle has not received nearly as much attention," he says.

Heat may also play a major role. Geochemists studying hydrothermal systems on volcanic islands off the coast of Italy have found that high water temperature can lead to higher concentrations of arsenic in the groundwater fed by those systems³.

Highway of the Pacific

Still, much is uncertain as to how the geology of a region influences the temperature and minerals that precipitate at different vent sites. The Galapagos spreading ridge, the site of Haymon's recent explorations, is one such example. A 'hot spot' of magma from deep in the Earth formed, and continues to form, the Galapagos Islands. The amount of magma and heat released from this hot spot has created a thick, possibly pliable crust in the region that could permit only low-temperature hydrothermal vents to emerge. And that's all previous expeditions had found.

But Haymon thought that the spreading

ridge north of the islands, where the hot magma lies close to the sea floor, would yield the focused flow of high-temperature fluid needed to form black-smoker chimneys. And nobody had yet explored this area for vents.

A few hours after discovering the first Galapagos smoker chimney, named Plumeria after a tropical flower⁴, Haymon's team spotted a cluster of as many as six black-smoker chimneys, 12 to 14 metres high. In a nod to the Galapagos' most famous marine reptiles, the team dubbed the chimneys the Iguanas. Two weeks later, the group came upon two other sites featuring more black smokers.

The discovery allows for the possibility that the underwater Galapagos chimneys, like the islands themselves, are home to unique animal species. Or it may be that the chimneys provide a highway for animals exploiting the special conditions the vents create; larvae could drift in the vent plume from chimney to chimney, up and down the mid-ocean ridge. "We saw animals, but I could not say from looking whether they were new species," she says. Only further exploration, collection of the organisms, and genetic testing will tell.

It may be fitting that the Galapagos — the site that fuelled Charles Darwin's ruminations on evolution — are now the focus of the latest underwater mystery. But Haymon isn't surprised. "When you consider the size of the ocean and how hidden everything is, we have only begun to really know what's there," she says. "I'm convinced there are many surprises left for us."

Christina Reed is a freelance science writer based in Seattle, Washington.



Iron-oxidizing bacteria living in and around deep-sea vents make conditions perfect for sea lilies.

1. Clarke, T. *Nature* **421**, 468–470 (2003).
2. Price, R. E. & Pilcher, T. *Chem. Geol.* **224**, 122–135 (2005).
3. Aiuppa, A. et al. *Chem. Geol.* doi:10.1016/j.chemgeo.2005.11.004 (2006).
4. Haymon, R. M. www.oceanexplorer.noaa.gov/explorations/05galapagos/logs/dec14/dec14.html (2005).



How green was my subsidy?

Europe pumps large quantities of cash into schemes that encourage less-intensive farming. But, finds **John Whitfield**, some researchers are not sure what benefits they deliver.

It's five-and-a-half hours on the train from London to Wageningen in the Netherlands. Through Kent and the Pas de Calais, the view is of rolling arable fields; north of Brussels, it changes to a flat checkerboard of pastures, edged by ditches and dotted with old windmills and modern wind turbines. These are some of the most intensively farmed landscapes on Earth, from which plants and animals unable to earn their agricultural keep have been excluded over many years.

In the time it takes to make this journey, about €2 million (US\$2.4 million) of public money will be spent, Europe-wide, on subsidies aimed at reversing the centuries-long trend towards the highest possible efficiency in farming. Together, governments and the European Union (EU) spend roughly €3.5 billion a year on schemes aimed at encouraging less-intensive farming in order to see gains in biodiversity, landscape preservation, and water and soil quality. These schemes cover a quarter of the farmland in the 15 countries that made up the EU before enlargement — an area about the size of Germany. But their continuation is not guaranteed. The latest EU budget deal could see funding for such 'agri-environment' schemes in these countries fall by a fifth compared with 2006.

As well as being a daunting spending commitment, Europe's agri-environment schemes also represent one of the world's biggest ecological experiments — or they would do, had anyone bothered to formulate hypotheses or collect data. Unfortunately, from this point of view, most of Europe's agri-environment schemes have very vague goals, such as to

"prevent damage to the environment" or "provide wildlife habitats". Specific targets are not set; progress is rarely monitored; the baselines from which they start are not defined. The good that they do is thus hard to measure, which in some eyes makes the schemes hard to justify. As the European Court of Auditors wrote in a critical report¹ on the subject last year, "If a measure cannot be adequately checked, it should not be the subject of public payment."

Going Dutch

At the end of last month, 75 researchers, as well as conservationists and officials from Spain, Estonia and most places in between, met in Wageningen to discuss measurements of the schemes. They looked at the largest international study of agri-environment strategies so far, and talked more generally about whether such schemes work, how to tell if they do, why some are successful, and how the rest could be improved. There is definitely a lot of room for such improvement. If judged in terms of biodiversity, research suggests that almost half the schemes have no effect. A handful have an effect opposite to that intended.

The meeting was convened by ecologist David Kleijn and his colleagues at Wageningen

University, who have pursued the subject with vigour. In 2001, in one of the first scientific audits of an agri-environment scheme, they showed that a Dutch project intended to help ground-nesting meadow birds by delaying the mowing of fields was having no effect — birds actually seemed to prefer intensively farmed fields². The paper caused a storm. The Dutch agriculture ministry, which did not take kindly to being told it was wasting its money, summoned the researchers to explain themselves. "They really didn't like it," recalls Frank Berendse, one of the team. "I was quite amazed by the impact," adds Kleijn.

Sue Armstrong-Brown, head of agricultural policy at Britain's Royal Society for the Protection of Birds, thinks that the Wageningen research contributed to the current cuts in agri-environment spending: "You can trace a lot back to the 2001 Kleijn paper, and the debate around it — that work caused a lot of concern among policy-makers."

Kleijn went on to win EU funding for an evaluation of agri-environment schemes across Europe, a project called EASY. He began by publishing a meta-analysis in 2003 of the evaluations up to then, working with William Sutherland of the University of East Anglia in Norwich, UK. They unearthed 62 studies, of which three-quarters had been conducted in Britain or the Netherlands; a third of the delegates at the Wageningen meeting were from Britain, reflecting the higher profile of nature conservation in the north of the continent relative to the south. Many studies were rejected owing to their poor design; of the rest, 54% showed some positive effect of schemes

B. WYNAR/ALAMY

"It was rather depressing, wading through report after report that showed minimal benefits."
— William Sutherland

on biodiversity, 6% showed a negative effect, and the rest no effect³. “It was rather depressing, wading through report after report that showed minimal benefits,” says Sutherland.

That same year, Kleijn and an international team of researchers began a study of schemes in the Netherlands, Spain, Germany, Britain and Switzerland, the results of which were unveiled at January’s meeting in Wageningen. This team chose pairs of nearby fields, in which one was part of a scheme and the other was not. They then surveyed the fields for five groups of creature: plants, bees, grasshoppers and crickets, spiders and birds. Out of 25 categories (five species groups in five countries), 12 showed a positive response to agri-environment schemes⁴.

Plants showed the most widespread benefits, with higher diversity on scheme fields in every country except the Netherlands. Bees benefited in Germany and Switzerland, grasshoppers and crickets in Britain, and spiders in Spain. In cases where the biodiversity went up, nearly all the beneficiaries were common species; only one scheme — a Spanish programme aimed at making arable fields bird-friendly by leaving winter stubble — showed a positive effect on endangered species, one of which was the thekla lark (*Gal-erida theklae*). “One could argue that this lack of benefits is the most important result, because endangered species are the ones that suffer most from the intensification of agriculture,” says Kleijn.

Green sleeves

His interpretation was not popular with everyone at the Wageningen meeting. British conservationists, for example, questioned the EASY team’s decision to study a British scheme that pays farmers to leave strips of grass at the edges of their arable fields, saying that it was never designed to protect rare species. “It would be a bit nuts to criticize a scheme for not delivering something it wasn’t designed to do,” says Juliet Vickery, head of terrestrial ecology at the British Trust for Ornithology in Norfolk.

Britain’s Environmental Stewardship programme of EU-approved agri-environment schemes has a small upper tier of focused, complex schemes aimed at preserving habitat for particular species, and a far broader lower

Verging territory:
farmers are often
paid to leave the
margins of fields
untouched.



R. PYWELL, CENTRE ECOL HYDROL.

tier of things such as grassy verges, which UK conservationists expect nearly all farmers to adopt. The programme has a strong research base, with monitoring planned for 2,000 sites. “It’s the best model for these schemes in the world,” says Armstrong-Brown.

“There’s a tremendous amount of expertise in Britain,” agrees John Finn, an ecologist with Teagasc, the Irish Agriculture and Food Development Authority, “but it isn’t yet being shared effectively with other countries.”

The model for the upper tier of these efforts is the campaign to improve the lot of the circl bunting (*Emberiza circl*). This English farmland bird was once widespread, but by the 1980s it had been restricted to a small part of the southwest — although it was still common in other countries. A sustained conservation effort has paid farmers to leave stubble for winter food and to preserve grassland as a source of summer insects. Together with intensive support from the Royal Society for the Protection of Birds, this has seen the population expand sixfold in less than two decades⁵.

To cap it all

Kleijn accepts that intensive conservation efforts can work, but by the end of the Wageningen conference he still feared that, without focus, few schemes would achieve tangible goals. Others, however, left the meeting worried that such an emphasis on measurable results could be counterproductive. “The EASY studies are very sound, but the take-home message tends to overstate the case,” says Simon Mortimer of the University of Reading, UK. “It could potentially be quite damaging to a policy that has had benefits.” Just because the benefits so far have been small, he adds, does not mean the schemes aren’t working.

And the failures can be understood by analysing what farmers do, and how they respond to incentives, says Jørgen Primdahl, an environmental scientist at the Royal Veterinary and Agricultural University in Frederiksberg, Denmark. Kleijn’s study would have been stronger had it done this, Primdahl maintains. “Just looking at the outcomes of schemes, without any reference to agricultural practices, is not

very helpful,” he says. “If you have good models, you can explain why a scheme doesn’t work.”

In the long term, the schemes may fare well despite a lack of clear data on their results. The public is more likely to approve of paying farmers to manage the landscape than to produce food surpluses, says Hans-Jörg Lutzeyer of the European Commission, one of only two commission officials to make it to Wageningen. While centralized EU funding for agri-environment schemes is being cut, member states have increased their freedom to switch money into environmental measures from other farm subsidies. World Trade Organization rules, which prohibit subsidies for food production but allow them for the environment, add to the attractions of agri-environment projects for Europe.

Such schemes may not be the best way to promote the preservation of endangered species. Sutherland, for one, argues that Europe might do better to allow some areas to revert to a state close to wilderness while others are intensively farmed, and then to manage the whole system so as to maximize leisure, flood protection, and water quality. He says biodiversity benefits would accrue even if not particularly targeted⁶. But Europeans like farmland landscapes, and will probably continue to try and convince themselves that there are practical ways to keep areas that are rich in wildlife and pleasing to the eye, which also produce cheap food and don’t pollute streams and rivers.

“It’s worth remembering how weird Western European conservation is,” Sutherland says. “If you tell American, Asian or African conservationists that we farm in national parks, they look at you in disbelief.”

John Whitfield is a freelance writer based in London.

1. European Court of Auditors *Official J. Eur. Union* Sp. Rep. No 3/2005 (2005).
2. Kleijn, D., Berendse, F., Smit, R. & Gilissen, N. *Nature* **413**, 723–725 (2001).
3. Klein, D. & Sutherland, W. J. *J. Appl. Ecol.* **40**, 947–969 (2003).
4. Kleijn, D. et al. *Ecol. Lett.* doi:10.1111/j.1461-0248.2005.00869.x (2006).
5. Peach, W. J., Lovett, L. J., Wotton, S. R. & Jeffs, C. *Biol. Conserv.* **101**, 361–373 (2001).
6. Sutherland, W. J. *Ibis* **146**, 230–238 (2004).

E. JONGEJANS



“The Dutch government did not take kindly to being told it was wasting money, and summoned David Kleijn to explain his results.”

BUSINESS

Rumblings from the fringe

The European Patent Office could fall victim of its own success, as threatened national offices attempt to claw back some influence. Alison Abbott reports.

These days, most universities and companies seeking patent protection for their ideas in Europe go to just one place — the sleek, imposing Munich headquarters of the European Patent Office (EPO).

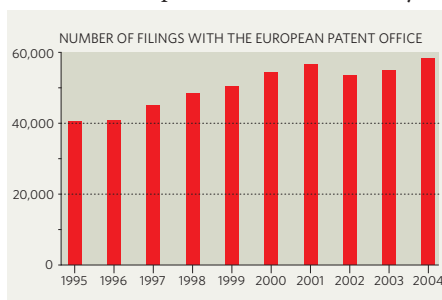
But behind the EPO's quiet facade, where almost 6,000 patent examiners process tens of thousands of patents each year, conflict is brewing. The 31 nations that have joined the office's founding convention no longer see eye-to-eye on the extent to which the European office should supplant national patent offices.

As is often the case with emerging European institutions, national interests look set to predominate next month, when the members of the convention meet in Munich to discuss whether some of the EPO's growing workload should be distributed to the national offices of its member states.

Two camps have emerged on the EPO's administrative council. One wants some of the EPO's work — currently done at Munich and subsidiary offices in the Hague and Berlin — to be devolved to national patent offices. The other wants to keep the work centralized.

"When the split emerged last year, it was very painful," says Ciaran McGinley, a principal director at the EPO. "It would be very wrong to dilute the density of expertise we have here."

The conflict pits France and Germany —



Office politics: the European Patent Office's ruling council is divided between devolvers and centralizers.

two of the original founders, whose languages are used, along with English, in all EPO documentation — against smaller nations, such as Finland, Sweden and Hungary, that don't want to let their own patent expertise wither away.

The smaller nations also fret that their inventors will suffer if they can't file European patents in their native languages. "There are 200 million citizens in EPO countries who have either English, French or German as their native language, and 300 million who don't," points out Eero Mantere, vice-president of the Finnish patent office. "We need equality."

For the researchers whose work leads to the roughly 60,000 patents filed to the EPO each year, this bureaucratic struggle may seem a distant concern. Some say their main worry is the waiting time that their applications face, as the EPO's backlog builds up (see graph).

But patent-office watchers are nervous that a move to spread the work could damage patent quality. "Why change a winning team?" says Christian Stein, chief executive of Ascenion, a Munich-based technology transfer organization in the life sciences that works for some of Germany's best publicly funded laboratories.

The 1973 European Patent Convention (EPC), which created the office, sought to save

money and improve quality by centralizing patent-granting at the EPO. It has substantially succeeded in that aim, and the office is now by far the favoured European venue for patent-seekers from all over the world.

EPC signatory countries share the cost of running the EPO and their representatives — who often work for national patent offices —

comprise the EPO's administrative council. But this arrangement can lead to conflicts of interest for council members, critics say.

And jostling between nations to control the EPO is never far below the surface. The council has been unable to choose which nation should hold the EPO's rotating six-year presidency, for example, and has instead split it between French incumbent Alain Pompidou and Alison Brimelow, the British patent official who will take over in 2007.

Partly, the struggle reflects the fact that some national offices no longer have enough to do. When a patent is filed, a literature search is conducted to find out if a similar invention already exists. Only then do the EPO's examiners try to establish if the application meets the criteria of patentability — novelty, inventiveness and utility.

At first, the EPO sent some preliminary

"Jostling between nations to control the patent office is never far below the surface"

searches to five national patent offices. But last year, in a move that helped trigger the current dispute, the EPO council decided that its examiners should do both the search and the examination on each application.

A second cause of the row was the Finnish patent office's decision to apply for Patent Cooperation Treaty status, to allow it to conduct searches for those considering filing patents to, for example, the US and Japanese offices. There is nothing to stop a European nation from doing this — but none had chosen to do so since the EPO was established in 1973.

Since then, all hell has been let loose inside the administrative council. Decentralizers want more of the action to be devolved, and maintain that their staff can readily match EPO quality.

On the edge

Lars Björklund, deputy director-general of the Swedish patent office, argues that countries on the geographical edge of Europe need to maintain their own expertise in intellectual property to support local inventors. "And we could help with the EPO's backlog," he says.

But Wim van der Eijk, EPO head of international relations, charges that those backing decentralization "are trying to reverse history". EPO officials claim that quality will suffer: McGinley says that the office would have liked to complain to the administrative council about the quality of searches carried out by two of the five national offices that had earlier shared EPO work. But it proved politically impossible to do so, because the relevant offices were represented on the council. "Everything quickly becomes politicized," he says. "You can put whatever quality control you want in place — you just can't enforce it."

The EPO is instead trying to win delegations over to a plan that it hopes will meet concerns that they might lose their expertise. Its proposed 'utilization programme' incorporates things such as training in all aspects of intellectual property, and building computer systems that provide national patent offices with full access to EPO documentation and software, such as that used to do searches.

The office is reconciled to negotiations beyond next month's meeting before the issue is resolved. In the meantime, users remain bemused. "I see no rationale other than political for splitting the office's work," says Stein.

London-based patent attorney Richard Jackson of Carpmals and Ransford, a computing specialist whose customers include University College London, says: "The quality of EPO work is good. I'd have thought that outsourcing would be the last thing it needs right now."

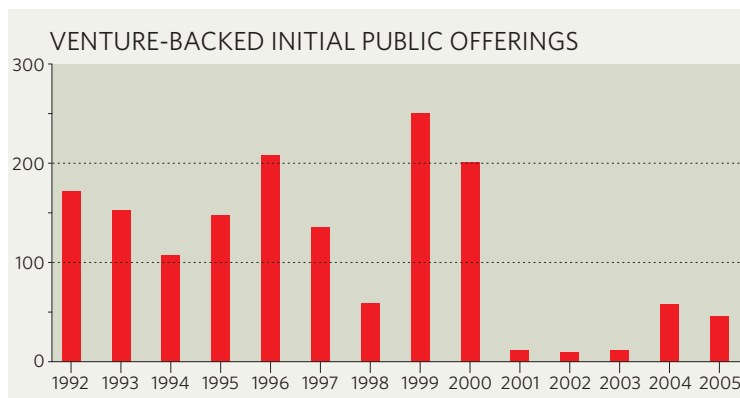
IN BRIEF

LIFT OFF OR RIP-OFF? Qinetiq, the UK technology company built from the research laboratories of the country's ministry of defence, successfully floated on the London stock exchange on 13 February (see *Nature* 436, 775; 2005). Shares in the company closed up 6% on the opening day's trading at £2.12 (US\$3.68), valuing the company at £1.3 billion. But the British government has been sharply criticized for its handling of the privatization, under which the Washington-based Carlyle group took a minority stake in the company in 2002 and fattened its value through the acquisition of several US defence contractors, before realizing an eight-fold return on its initial investment of £42 million.

SUCCESSION STRUGGLE SkyePharma's board of directors has ousted the company's founder and chairman, in an attempt to clear the way for the man it earlier this month appointed to succeed him. The board installed Jerry Karabelas, a former chief executive of Novartis' drug division, to replace Ian Gowrie-Smith. However, it must win shareholders' approval of the appointment at an extraordinary general meeting that will be held in the next few weeks. Dissident shareholders in the London-based drug delivery firm are backing Bob Tian, current head of laboratory-supplies company Whatman, for the job.

AVASTIN SETBACK Enrollment in a clinical trial of Genentech's blockbuster cancer drug, Avastin, has been suspended because of non-cancer-related deaths in one group of patients taking the drug. The company said that seven people with early-stage colon cancer, some of them young, died in a group that was taking Avastin in addition to standard chemotherapy; two of these patients died of cardiac causes. In a group of patients on standard chemotherapy alone, four died from non-cancer causes. Sales of Avastin in the United States last year were \$1.1 billion.

MARKET WATCH



The number of initial public offerings for US companies that had been backed by venture capital fell back during 2005, and shows little sign of returning to its 1990s levels.

Low investor interest, together with regulatory changes on Wall Street that have made it more complicated and expensive for companies to obtain a public listing, have made it more likely that venture capitalists will get their money back by selling young companies directly to large corporations, analysts say.

The number of public offerings fell from 67 in 2004 to just 41 last year, according to data compiled by research firm VentureOne, a San Francisco-based unit of Dow Jones. This is sharply down, not just on the days of the dot.com boom, but also on the regular

level of such offerings in the years before then (see graph).

VentureOne research analyst Josh Grove points out that corporations have no shortage of cash available to make such acquisitions. Last year, they forked out more than \$27 billion to make them happen, up almost one-fifth on 2004. "They're snapping up these companies before they can go public," he says.

Thirteen of last year's offerings were biotechnology firms and 11 specialized in information technology. In biotechnology, investors have become particularly wary of buying into companies until they are running late-stage clinical trials. "The demand is not there," says Michael Lytton, a general partner of Oxford Bioscience Partners, a life-sciences venture capital company based in Boston.

Helen Pearson

Public repositories need serious funding

SIR — We support the suggestion made by Carlos Santos and colleagues in Correspondence (*Nature* **438**, 738; 2005) that data associated with peer-reviewed articles should be submitted to recognized, public repositories wherever possible.

We suggest that attaining this goal requires the support of national and international funding bodies that are willing both to implement data policies and to fund efforts to create community-driven standards and public repositories.

For funding bodies, supporting a data policy can be expensive. Current policy for the UK Natural Environment Research Council (NERC; www.nerc.ac.uk) states that all data generated by projects funded under the environmental-genomics programme and the post-genomics and proteomics programme must be submitted to a suitable public repository, when one is available. NERC puts approximately 12% of the funds from each of these programmes towards data management and training through the establishment of the NERC Environmental Bioinformatics Centre (<http://nebc.nox.ac.uk>), which facilitates 'omic' data management by developing data standards, software, databases, bioinformatics workstations and courses, and delivering these to the community, as well as hosting digital data for cases where suitable public repositories do not already exist.

We argue that putting the tools and facilities in place to enable good data management is an area worth investing in. As well as addressing the aims of integrated, long-term data storage and access, this investment would minimize duplication of effort, facilitate uptake and sharing of data and maximize the potential for comparative analyses.

Bela Tiwari*, Dawn Field*, Jason Snape†

*NERC Environmental Bioinformatics Centre, CEH Oxford, Mansfield Road, Oxford OX1 3SR, UK

†AstraZeneca UK Ltd, Freshwater Quarry, Brixham TQ5 8BA, UK

Public repositories: users reluctant to give materials

SIR — We welcome your call to debate whether journals should impose more stringent publication criteria for biological materials, notably cell lines ("Standards for papers on cloning" *Nature* **439**, 243; 2006). Our mission at the DSMZ cell repository, the German national resource centre for biological material (www.dsmz.de) is to provide investigators with authentic, well-characterized biological material. But it is all

too often compromised by users' reluctance to make their own materials freely available. About 90% of researchers contacted since 1994 after claiming to have established newly published cell lines have refused or simply ignored our requests for material. Only one in 580 contacts indicated a preference to deposit cells with another facility, and a search of a large sample of cell lines currently held by other major repositories revealed none previously denied to the DSMZ.

Even among the minority depositing human tumour-cell lines at the DSMZ 29% provided identifiably false, cross-contaminated cell lines (CCCL). Although reliable detection of CCCL is limited to institutions such as ours that hold cell lines en masse and DNA-profile all accessions, this is almost certainly an underestimate of the true level. The systematic use of CCCL is likely to generate distorted data: accordingly, the Sanger Institute has highlighted instances of CCCL among the NCI-60 reference cell line panel used to identify tumour-specific gene expression (www.sanger.ac.uk/genetics/CGP/NCI60). But if new cell lines are not deposited, they cannot be authenticated.

Initial descriptive publication is the only effective stage at which sanctions might be applied. Journals such as *Nature* and *Science* should adopt rigorous criteria requiring deposition of new cell lines, and encourage specialist journals to follow suit. Publications using existing cell lines should identify where these came from. Funding bodies could also insist that cell lines established with their support be submitted to public repositories. And those in charge of large collections could do more to clarify the availability and authenticity of their cell lines.

Roderick A. F. MacLeod, Hans G. Drexler

DSMZ — German Collection of Microorganisms and Cell Cultures, Mascheroder Weg 1b, 38124 Braunschweig, Germany

Diet's healthy blend of science and practicality

SIR — Your Editorial "A recipe for trouble" (*Nature* **438**, 1052; 2005) raises concerns with me of a disconcerting mindset around the 'preciousness' of science, which I sense exists in parts of the scientific community. The view that working with industry demeans the quality or integrity of science, or that using the media to translate scientific findings into uptake strategies is tainted by commercialism, is worrying. If we continue to give the perception that science is the stuff of ivory towers, not everyday life, then we will see ever greater reductions in research funding over the years to come.

By working in partnership with industry, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) created the

Total Wellbeing diet, which can contribute to reducing obesity in Australia. This has been an outstanding success for the CSIRO, not because it is returning funds to the organization for further research, but because it represents a successful translation of research findings to the public in a way that can be understood and absorbed into everyday lives. Just as the CSIRO was pleased to have its name associated with the peer-reviewed publications associated with the diet, so it continues to support further communication of the work to the public through the diet book. It has not been manipulated either by the food industry or by the publisher.

Of course, human nutrition is complex, and a lively debate continues about the benefits of high-protein diets (see, for example, www.atkinsdietalert.org/advisory.html). But the book's high-protein, moderate-carbohydrate eating plan is based on peer-reviewed science within robust experimental frameworks and the term 'scientifically proven' is justified.

Alastair Robertson

Group Executive, Agribusiness, CSIRO, PO Box 93, North Ryde, New South Wales 1670, Australia

What goes around comes around in drug discovery

SIR — We read with interest your Business story "Merck opts for shake-up to clear drug pipeline" (*Nature* **438**, 1076–1077; 2005), which concluded with the thought-provoking quote by an industry analyst: "Merck is good for the pharmaceutical industry. You can't make me-too drugs unless you have someone to copy."

We were drawn to this statement since your story focuses on Merck's anti-nausea drug Emend, without pointing out that Emend itself could be considered a drug based on incremental scientific advance. Merck scientists Malcolm MacCoss and Thomas Baillie, discussing the history of Emend (*Science* **303**, 1810–1813; 2004), trace its direct lineage to the publication of the discovery of non-peptidic substance P antagonists (R. Snider *et al.* *Science* **251**, 435–437; 1991).

Our view, as three of the co-authors of the 1991 paper, is that the world needs a healthy and innovative pharmaceutical industry and that, as such, we are all good for each other.

Robin Spencer, Kelly Longo, John Lowe

Pfizer Global Research and Development, Eastern Point Road, Groton, Connecticut 06340, USA

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. They should be signed by no more than three authors; preferably by one. Published contributions are edited.

BOOKS & ARTS

The changing face of Arab culture

Studies of the differences between people can shed light on the rise of Islamism in the Middle East.

The Middle East: A Cultural Psychology

by Gary S. Gregg

Oxford University Press: 2005. 472 pp.
\$45, £26.99

Robert Springborg

It makes one feel like an intellectual fossil to encounter a book that harks back to a subdiscipline that is now virtually extinct, but which was still in vogue when one was a graduate student. Cultural psychology flourished in the wake of the Second World War as a result of desire to understand totalitarianism. Following in the intellectual footsteps of anthropologists such as Margaret Mead and Ruth Benedict, scholars interested in the Middle East and North Africa sought to document and explain the details of the region's cultures and what they viewed as its distinctive personality types. Daniel Lerner's landmark study, *The Passing of Traditional Society* (Macmillan, 1958), was the definitive statement of the cultural, personal and developmental consequences of the clash between tradition and modernity in these regions.

But by the late 1960s, such works were criticized for their ethnocentrism and for having ignored global and national systems of economic and political power. At the same time, the appearance of studies that seemed hostile to Arab Muslims left the field of cultural psychology, at least in its Middle East variant, open to accusations of racism. Edward Said's *Orientalism* (Routledge, 1978) administered the *coup de grâce*. Except for some interesting investigations by Arab scholars, such as Hisham Sharabi and Halim Barakat, who sought to understand patriarchy and other aspects of cultural psychology that they saw as characteristic of the region, the field largely died out.

The spread of rational-choice theory from economics into other social sciences ensured that the disappearance of cultural psychology was not lamented. It is the sameness of people, regardless of their culture, that now dominates. We are all global citizens driven by desires to maximize various utility schedules, with variation in those schedules being the only possible area within which culture might be attributed status as an independent variable. The human, in short, was largely written out of studies of development, a field which itself began to disappear in the 1980s as even



The Bibliotheca Alexandrina in Egypt highlights the region's mix of the traditional and the modern.

systemic differences between the developed and developing worlds were deemed irrelevant.

It took Islamist-inspired political violence, especially suicide bombing, to rekindle interest in the particularities of the Middle East and North Africa. For some years explanations have been sought in the theology and practice of Islam, in the consequences of the region's failed economic and political development, and in the impact of Western policies on the region. It is only now, however, that we have the first major study that draws on cultural psychology to address the issue of whether the cultures and individuals of the region have identifiable particularities, and, if so, whether they may account, at least in part, for Islamism and other sociopolitical phenomena.

This is a much-needed and noble effort. Resuscitating a moribund subdiscipline that is based on the premise of difference rather than similarity between peoples is important, not least for serving as a check on other social-science approaches. We need to see if human difference can serve as the foundation for rewarding intellectual enquiry, rather than descending into crude national stereotyping. Gary Gregg claims in *The Middle East* to be aware of this trap, but some readers may feel he has fallen into it. While I am neither so critical nor so dismissive, I do have reservations about

the structure of his argument and the evidence presented to support it.

Taking the issue of evidence first, the overall impression is that it may be insufficiently representative. Gregg's fieldwork experience in rural Morocco seems to have tilted his perspective towards that country and rural areas in general. This tendency is reinforced by his reliance on anthropological work, which has also been conducted primarily in rural areas. Yet the Middle East and North Africa have urbanized more rapidly than any other areas of the globe, and most of their citizens now live in cities.

Gregg also draws heavily on characters in the novels of Egyptian author Naguib Mahfouz, for example, to provide illustrations of stages of personal development. That Mahfouz's characters are often overdrawn and clearly created to personify idealized types is but one reservation about the appropriateness of using such material to elucidate a particular psychological model. It is disappointing that Gregg has not included evidence that is more relevant. Pioneering psychologists and psychiatrists began plying their trades in the Middle East and North Africa in the late 1940s, so we are now into the second and even third generations of these professionals. They could provide a veritable treasure trove of case-study

material. It might also be instructive to know how they have theorized about personalities and how those theories might differ from those of practitioners in other settings.

A central proposition of Gregg's book is that discontinuities and outright failures in the region's economic and political development have had negative effects on the psyches of large numbers of its inhabitants, causing them to turn to fundamentalist Islam. But this assertion also goes largely unsupported by directly relevant data — even though the case histories of young Arabs, including those who have turned to religion and violence, are accessible and have been drawn on by other authors interested in the link between personal circumstances and political violence.

The structure of Gregg's argument seems questionable in at least two important regards. First, his contention that the Middle East and North Africa constitute a coherent cultural area can be challenged. He admits that there are subregions, including North Africa, the Gulf and the Nile Valley, but he de-emphasizes the differences between them. The absence of the Levant as a distinct subregion is one shortcoming. The more vital question of whether

a Beirut and a dweller in the Moroccan High Atlas, for example, inhabit the same cultural space is largely unanswered.

A second problem has to do with his use of the dichotomy between the traditional and the modern. He repeatedly warns that this duality is an oversimplification, but much of his argument is nevertheless based on discontinuities resulting from people having to inhabit these two divergent worlds, or moving from one to the other. His argument about the region being a distinct cultural area is based heavily on the legacy of traditional patterns of agriculture and desert nomadism, as if the region had not changed since Lerner wrote his seminal work in the 1950s.

But the region has changed dramatically since that time, raising the question of the extent to which the traditional continues to clash with the modern and thus accounts for personally disruptive discontinuities. Gregg states that young high-school and university students typically have illiterate parents, as if the region were witnessing its first generational wave of secondary and post-secondary education. In reality, many countries in the region are now into their third, fourth or even

fifth generation of university-educated youth. The connection between today's youths and their camel-herding forbears, if indeed that's what they were, is pretty tenuous. Cultural psychology should long ago have abandoned this hoary old duality and come up with explanations of contemporary cultures and the personalities of which they are constituted that rely more on the present than on some hypothetical past.

Despite these frailties, Gregg's book has much to recommend it. The publishers are to be commended for producing a book in a discipline widely considered to be passé, but which deserves reconsideration. Gregg provides an encyclopaedic review of the cultural-psychology literature while seeking to reintroduce individual variation as an important variable for understanding the key issues of our times. He raises the important question of the impact of oppression, war and violence on large numbers of residents of the Middle East and North Africa, and provides a useful agenda for future research. ■

Robert Springborg is at the School of Oriental and African Studies, University of London, Thornhaugh Street, London WC1H 0XG, UK.

Feathered friends

In the Company of Crows and Ravens

by John R. Marzluff & Tony Angell

Yale University Press: 2005. 408 pp.

\$30, £18.95

Crows: Encounters with the Wise Guys of the Avian World

by Candace Savage

Greystone Books: 2005. 120 pp.

\$20, Can\$27

Alex Kacelnik

Konrad Lorenz claimed that a tame crow called Hansl, returning after a long absence with a broken digit, said the German equivalent of "Got him in a blooming trap!", and so, by repeating the words of its captor, informed Lorenz how the injury had occurred. Today we would see this as an unjustified claim of the use of referential language and declarative autobiographical memory. This striking example of projective anthropomorphism is not unusual when dealing with crows, either in popular culture or in the scientific literature. People who live around crows, ravens or other corvids often see them as exceptional among animals, possessing qualities of cunning, reasoning, deception and, frequently, magic.

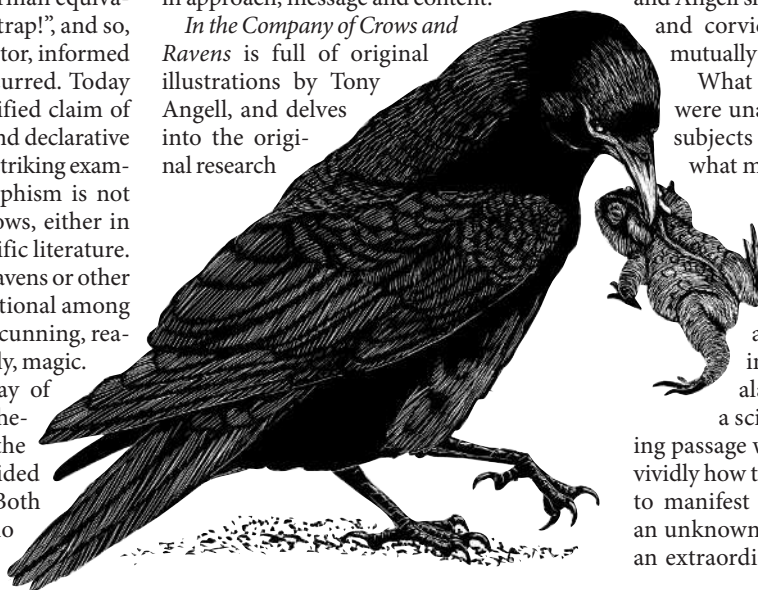
I can think of no better way of becoming immersed in this phenomenon — and enjoying the experience — than that provided by these two beautiful books. Both have been written by people who know and love these animals,

and who use every opportunity to introduce, in lucid but accessible language, up-to-date ideas from behavioural ecology and evolutionary biology. Both are beautifully illustrated and produced, and this adds to the pleasure of sailing easily through them. Indeed, they are as gripping and difficult to put down as any good work of fiction. They mix scientific research with fables, poems and mythical stories ranging from Scandinavia to Mexico and Australia. The ground they cover overlaps to some extent, but there is sufficient extra material in each to make it worthwhile to buy both, as they differ in approach, message and content.

In the Company of Crows and Ravens is full of original illustrations by Tony Angell, and delves into the original research

published over many years by John Marzluff. Marzluff and Angell put forward the notion of cultural coevolution, and they suggest that corvids are the clearest example of a culturally coevolved wild organism. They even hypothesize that the very existence of human culture as we know it may have been influenced by corvids. They write, for instance: "Ravens scavenged from large animal kills in the Pleistocene and quickly learned to exploit the foods gathered by early human hunters and fishers. Fending off scavengers may have favored people with a culture of living in groups." By referring to many published studies, ranging from these ancient interactions to the present-day association between crows and the agricultural and urban environment, Marzluff and Angell show how the destinies of humans and corvids have been intertwined and mutually influential.

What a pity, then, that the authors were unable to resist the charm of their subjects and have on occasion fallen to what might be called 'corvidian exceptionalism'. Not only are some anecdotal observations used to generalize concepts beyond what can reasonably be justified, but, in a few cases, the authors seem to have indulged in a suspension of disbelief that is alarming in a book popularizing a scientific matter. In an embarrassing passage we are told: "Tony Angell recalls vividly how the spirit of a good friend seemed to manifest itself in a crow." It seems that an unknown crow appeared and behaved in an extraordinary fashion for two days until



Angell realized the crow's similarity to a friend who lived far away and had been ill. Needless to say, Angell later learned that his friend had died the day the crow arrived. They then ask people to report similar experiences. Do not be discouraged, however, as this is the worst such transgression in the book, and it should not obscure the value of the rest. Perhaps you should skip that page.

Candace Savage's *Crows* is a beautifully crafted celebration of these birds, and places a greater emphasis on the author's and other

people's artistic and emotional perception than on the fostering of a specific hypothesis. But she too uses every opportunity to introduce sound evolutionary and ecological concepts, including good descriptions of recent research on crow communication, breeding biology and, to a lesser extent, ecology.

Corvids are without doubt extremely interesting creatures, and research on corvid behaviour is enjoying an unprecedented boom. This is to be celebrated because the mapping of minds that are so different to the better-known

minds of mammals may help us think more clearly about the conditions that promote the evolution of advanced cognition. Corvid research may ultimately help in explaining why the primates, which perhaps deserve to be called featherless corvids, have experienced such rapid and differentiated cognitive evolution. Whatever you read, however, remember that crows are just birds, after all. ■

Alex Kacelnik is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

In the beginning

Genesis: The Scientific Quest for Life's Origins

by Robert M. Hazen

Joseph Henry Press: 2005. 339 pp. \$27.95

Leslie Orgel

There aren't many facts or opinions about the origin of life that are universally accepted. Researchers agree that Earth is a little more than 4.5 billion years old. Most would also agree that organisms more or less like bacteria evolved in the first billion or so years of Earth's history, and that darwinian selection, acting on polymeric molecules, was involved. But almost everything else about the origin of life remains obscure. Little is known with certainty about the physical environment in which life evolved or about the detailed steps that led from unconstrained abiotic chemistry to the organized complexity of biochemistry.

Our limited knowledge of the environment in which life evolved has led to the blossoming of diverse 'scenarios' aimed at describing one or more of the important steps on the road to life. Some are supported by extensive experimental studies, whereas others remain purely hypothetical. This is chaotic intellectual territory, so anyone entering it for the first time needs an accomplished and unbiased guide. Having read Robert Hazen's book *Genesis*, I can recommend him strongly for the task.

One problem facing authors of introductory books on the origin of life is the popularity of two almost antagonistic scenarios. Stanley Miller's classic experiments on the synthesis of amino acids in a reducing atmosphere led to the idea of a 'prebiotic soup', according to which a genetic system evolved directly in the surface waters of the early Earth. In contrast, Günter Wächtershäuser argued that a complex, non-genetic metabolism evolved on the surface of metal sulphides in geothermal environments, probably in deep-sea vents, before the appearance of a genetic system. Hazen is a metabolist, but he is even-handed in his treatment of the two approaches, at times making it clear that one of his opinions is a minority view.

Like a good guide, Hazen visits all the major



A difficult start: quite how the barren early Earth gave rise to life remains shrouded in mystery.

sites. Naturally, Miller's experiments are on the itinerary, but visits to all the prominent experimental programmes follow. Hazen is sympathetic to theoretical models and includes lengthy descriptions of some highly speculative proposals. However, he always makes it clear that these proposals are speculative and lack experimental support.

Hazen's style is lively and he has a knack for describing clearly and concisely the way that experimental techniques, such as mass spectrometry, work. His peripheral material is well chosen, if not always directly related to the origin of life. I was surprised to learn that if I was reconstituted with all of my ^{12}C replaced by the ^{13}C isotope, my weight would increase by three pounds. And I was fascinated by his account of a memorable confrontation between experts on Precambrian fossils. Diversions of this kind reduce the fatigue of the journey.

I particularly liked the biographical and autobiographical sections, as it is good to know something about your guide and his friends. Hazen is generous in his brief descriptions of

his collaborators' personalities. His accounts of his own experiences in the lab give a vivid picture of what it is like to be an experimentalist. You can imagine him watching intently as the pen of a recorder approaches a critical point, and willing a peak to appear.

The text is accompanied by copious notes and a list of well-chosen references. Unfortunately, there is no indication in the text of their existence. To find the notes you must go to the end of the book, where they are organized by text page, and the references can only be accessed from the notes. This inconvenient arrangement could perhaps be addressed in a later edition.

But this criticism is minor. This book provides the best overview of the 'origin of life' field for the non-specialist reader that I have encountered. I think that even those who are familiar with most of its contents will enjoy the presentation. ■

Leslie Orgel is at the Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA.

Bauhaus at the zoo

Modernist designers in the 1930s found inspiration in the life sciences.



BRITAINVIEW.COM

The penguin pool at London Zoo was built in the Bauhaus style to show how animals can thrive in novel environments.

Peder Anker

In the mid 1930s, faculty members from the Bauhaus design school in Germany fled Nazi harassment and moved to London. Among them were the school's former director, Walter Gropius, and the head of its metal workshop, László Moholy-Nagy, whose art will be on display next month at the Tate Modern in London. The radical designers, who sought to unify arts and crafts with industrial universalism, found new allies and patrons in a community of biologists in London who adapted Bauhaus architecture and art as part of their scientific vision for the future.

One of these was the ecologist Julian Huxley. As secretary of the Zoological Society of London he had an apartment at the zoo, which he used partly as a showroom for modernist design. Here, scientists, artists, architects, environmentalists and the science-fiction writer H. G. Wells regularly met for discussions about how to save humankind from environmental, economic and social destruction.

Bauhaus design was one of the group's chief passions, and Gropius looked to Huxley and his friends with hope and admiration. Traditional architecture and design reinforced an unfortunate dualism between people and nature, Huxley believed, whereas the Bauhaus approach promised a harmonious reunion. To Huxley, nothing less than the evolutionary survival of the human species was at stake.

The zoo became the testing ground for these architectural ideas. The zoologist Solly Zuckerman believed that the difference

between humans and animals was "almost certainly one of degree only", and he saw the life of primates as "a crude picture of a social level from which emerged our earliest human ancestors".

Visitors to the zoo could observe their own primitive desires in animals, Zuckerman argued, so it was of moral importance to place the animals in a model home for healthy living. The gorilla house and the penguin pool, along with a series of other buildings, were therefore built in the Bauhaus style.

But it was also politically important to the group to display thriving animals such as penguins in a highly unnatural setting, to show that humans too could prosper in new environments. "The most unlikely animals seem to thrive under what would seem the most unnatural conditions," zoologist Peter Chalmers Mitchell observed, if they have "freedom from enemies, regular food and general hygiene". The same would hold for workers and the poor, who desperately needed to be liberated from their 'natural' condition of criminal and filthy slums.

Moholy-Nagy, meanwhile, had found a major source of inspiration in Raoul Francé, a biologist and author who argued that humans should learn to copy nature's own inventions. Francé founded the science of bionics to pursue this end, calling it 'bio-technique'. Moholy-Nagy followed suit and told designers to use nature as a "constructional model" and to look for "prototypes in nature" to determine functional design. He read the work of Huxley and his friends and used it to generate his own principles, techniques

and processes that could be applied to human design.

Moholy-Nagy's 1935 film *In the Cradle of the Deep* documents the growth of lobsters and the fisherman's struggle to search them out. He argued that a "prehistoric animal shell is constructed in such a wonderful way that we could immediately adapt it to a fine bakelite or other moulded plastic form". The point of the film was to show designers that observing the life of animals can teach us about how form follows function.

His investigations into architecture and photography were also informed by the life sciences. He defined architecture "as an organic component of living" and argued that "architecture will be brought to its fullest realization only when the deepest knowledge of human life in the biological whole is available."

It was while living in London that Moholy-Nagy revolutionized the art of photography by capturing forces of life in action. By taking snapshots (as opposed to posed photos), he sought to replace static classical art with dynamic images of active people. He also tried to highlight the activity of places and objects. He used light and shade in a photograph of a laboratory tube, for example, to evoke a sense of gas moving within the glass pipe.

His artwork can be seen alongside that of fellow modernist Josef Albers in the exhibition "Albers and Moholy-Nagy: From the Bauhaus to the New World" at Tate Modern in London from 9 March to 4 June. Peder Anker is a research fellow in history at the University of Oslo, 0315 Oslo, Norway.

Family values in black and white

The sexual behaviour of penguins, although fascinating, does not offer the moral lessons that some popular commentators would have us believe.

Marlene Zuk

Anyone who doesn't realize that the intimate details of penguin family life are a hot political topic must have spent the last couple of years on an ice floe. First was the proliferation of same-sex penguin couples, most notably in Manhattan's Central Park Zoo. Roy and Silo, male chinstrap penguins, attracted international attention when they paid no attention to females in their enclosure, instead building a nest together and attempting to incubate a rock. After a sympathetic keeper gave the pair an egg, a chick called Tango emerged, who became the star of a children's book *And Tango Makes Three*.

The book — and the story — was embraced by gay-rights activists, who claim it illustrates the normalcy of homosexual families, an assertion hotly contested by conservatives. When, in a dramatic turn, the dapper couple split, with Silo taking up with Scrappy, a female from (where else?) California, right-wing website WorldNet-Daily gloated over the break-up. The site predicted that the subject would be "dropped from the media 'A list' faster than you can swallow a smelt".

On the contrary, last summer, the French documentary *March of the Penguins* became an unexpected box-office hit, and the penguin wars heated up. The film depicts the travails of emperor penguins as they migrate in subzero temperatures from the Antarctic ocean to their breeding grounds 70 miles away, mate, raise their impossibly adorable chicks, and return to sea. Perhaps because father penguins take a major role in child-rearing, spending weeks without food while guarding the egg and chick, the film also became a flashpoint for debates about family values. Conservative columnists praised the penguins' fidelity and sacrifice. Some went further and claimed the birds' behaviour upheld ideas about 'intelligent design'. Meanwhile, liberals scoffed at the idea that the penguins' lives of arduous deprivation represented the work of a creator; columnist Ellen Goodman observed, "if that is intelligent design, the big guy has quite the sense of humour".

Yet in the zeal for snappy headlines, some important points were overlooked. First, Roy and Silo are hardly the only same-sex couple either among penguins or in the

animal kingdom; scientists are increasingly recognizing such behaviour in a wide range of species, from sheep to gulls. Its significance lies not in its ability to answer the question of whether homosexuality is natural, much less whether gay marriage or adoption should be legal, but in the insight it offers into what sex means. Those unfamiliar with the lives of animals in the wild often assume that sex occurs in a brisk businesslike fashion, solely for procreation. In fact, sex often achieves much more; in



Penguin behaviour is fuelling debates about monogamy.

bonobos, highly social relatives of the chimpanzee, sex can defuse tense situations, and it occurs between members of the same sex, between adults and juveniles, and when conception is highly unlikely.

A far more interesting question than whether Roy and Silo vindicated the right or left wing is why we see pairings such as theirs in some species and not in others. Why in penguins and not in peacocks? Selection for pair bonds between males and females is likely to be the key starting point. Penguins as a group are monogamous, because it takes both parents to rear the chick, so the evolutionary seeds for seeking out a single mate are far more likely to take root in them than in the flamboyant peafowl, where a single mother can raise a brood alone.

Moreover, mate fidelity varies among penguins, with the emperors, rather ironically, showing high rates of partner switching between seasons (something the film notes but religious commentators conveniently ignore). Scientists speculate that the environmental constraints on breeding are so severe for the emperor penguins that they cannot afford the luxury of searching out the same partner they had the year before.

Conservative journalist Warren Throckmorton commented, "About the only thing

we can say from the Roy, Silo and Scrappy love triangle is that sexuality in animals is flexible, context-driven and influenced by factors we do not fully understand". But this is actually a rather profound conclusion, suggesting that animal sex could be far more complicated than a quick instinct-driven coupling. In other words, animal sex can be a lot like human sex, which may not be what Throckmorton intended to say.

Tom Turnipseed, writing for the website Zmag.org, suggested that the real message lies in the penguins' "cooperating with one another and sacrificing their own lives and individual gain for the common good and survival of their own kind" — behaviour that executives at Enron, the US energy company involved in an infamous corruption scandal, should have emulated. Other reviews also allude to this supposedly altruistic behaviour and the "inexplicable love" shown.

Were we watching the same film? In fact, the penguins are perfect little darwinians, selfish as can be. No one seemed to question why the birds took such pains on their return to the breeding grounds to find their own mate, their own chick, in a crowd of thousands of look-alikes. It seemed human, after all, like sailors returning from war eagerly seeking their families among the throng on shore.

But if the penguins simply needed to save the species, surely any chick would do, and feeding the nearest hungry beak would save all that tramping through the snow searching for one's special little one. Why bother? Evolution supplies the answer: only scrupulous discrimination of your own kin will perpetuate one's genes. How the penguins manage such sophisticated feats is a fascinating area of study, one that will yield much more than a consideration of whether they are good role models for monogamy.

If we use animals as poster children for ideology, we not only end up in meaningless arguments over whose examples are more significant (cannibalistic mantids or promiscuous bonobos?), we risk losing sight of what is truly interesting and important about their behaviour. What the executives at Enron are supposed to learn is another story.

Marlene Zuk is in the Department of Biology, University of California, Riverside, California 92521, USA.

© WARNER BROS. PICTURES 2005

ESSAY

NEWS & VIEWS

QUANTUM INFORMATION

To compute or not to compute?

Jonathan P. Dowling

Quantum physics aims another blow at common sense: a simple quantum computer gives the right answer, even when it is not run. (Traditionalists be comforted: the computer must be turned on.)

Is it possible to get a sensible answer from a computer without even running a program? Yes indeed, report Hosten *et al.* on page 949 of this issue¹ — and they have the experimental proof.

The thread that leads thus far has its origin in the early 1990s, with the introduction of a quantum paradox known as the Elitzur–Vaidman bomb thought-experiment². They showed that by sending a photon into a simple optical interferometer (Box 1), one can sometimes detect the presence of a light-triggered bomb without setting it off — in other words, without light interacting with it at all. ‘Sometimes’ — there’s the rub. At other times, you set the bomb off; at still other times, you get no information about the existence of the bomb at all. But then that’s the probabilistic nature of quantum mechanics. It might seem to defy common sense, but the result is a straightforward application of the quantum theory of light.

A few years ago, it was pointed out³ that an all-optical quantum computer could replace Elitzur and Vaidman’s bomb. Rather than exploding or not, this computer either runs a program or doesn’t, depending on whether or not a single photon impinges on its ‘run’ switch³. And just as we can obtain the answer to the question ‘is the bomb there?’ without exploding it, so we can obtain the right answer from the computer without running it. (The computer must, at least, still be turned on and properly programmed in order for the scheme to work.) However, although intriguing, this original ‘counterfactual computation’ scheme suffered from the same quantum-probabilistic issues as the bomb business: sometimes the device would actually run the quantum computer to get the answer; sometimes it gave no answer at all. The probability that it would give the correct answer while not running the computer was so small that one could always do better by guessing the answer randomly. Not much of a computer, that.

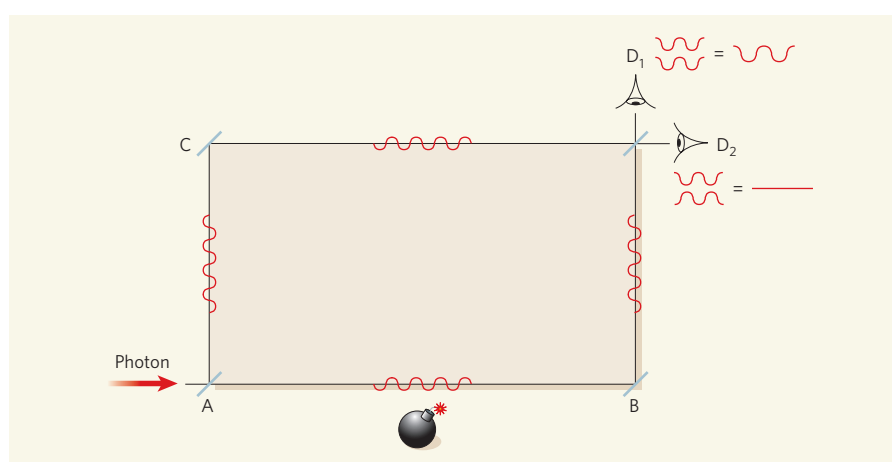
Another spin-off of the bomb idea was Anton Zeilinger and colleagues’ elegant and real (as opposed to thought) experiment called ‘quantum seeing in the dark’^{4,5}. That experiment allowed safe, non-explosive objects such

as a human hair to be imaged, sometimes using no photons at all. Again, the key word is ‘sometimes’: just as with the bomb, sometimes no photons were needed to image the hair; sometimes some photons were needed to image the hair; and sometimes the hair was not imaged at all. By sorting the individual imaging events carefully, however, an image of the hair could be constructed using only the events in which no photon had interacted with the hair. Better still, the probability of these photon-free imaging events could be increased using yet another paradox of quantum mechanics, the quantum Zeno effect⁶.

This effect is named in honour of Zeno of

Elea, the philosopher who gave us paradoxes such as that of Achilles and the tortoise (you can run, but you can’t catch up). It is the quantum-physical equivalent of the saying “a watched pot never boils”. If you measure the state of a quantum system often enough and fast enough, the system will remain in that state and never evolve to another state — even if it would evolve if you were not measuring it. Zeilinger and colleagues applied this idea by using a sequence of polarization rotators as ‘Zeno boosters’ that forced their interferometer system to remain far more often in the state of ‘imaging the hair with no photons’, and far less often in the states of ‘imaging the hair

Box 1 How to stop worrying and learn to love the bomb



Avshalom Elitzur and Lev Vaidman’s bomb thought-experiment² uses a simple wave-splitting device known as an optical interferometer and the quantum phenomenon of wave-particle duality. If just one photon enters their interferometer at A, its associated wave splits, going with a certain probability by way of either B or C. The geometry of the interferometer is set up such that, at one of two detectors at its far end (D_1), the waves from the two paths interfere constructively, so light is detected. At the other detector (D_2), on the other hand, the waves interfere destructively, so it is never triggered — unless one of the pathways, say B, is blocked off, preventing the passage of a photon. In this case, there can be no interference,

so the photon will be detected with equal probability by D_1 or D_2 .

Now suppose that B is — possibly — blocked by a photon-triggered bomb. A single photon enters the interferometer. Three outcomes are possible. First, the bomb goes off: so there is a bomb. Second, D_1 triggers. This outcome does not help us, as it could mean one of two things: there is a bomb, but the photon passed by C, or there isn’t a bomb, and waves from the two paths are interfering constructively. Third, D_2 triggers. In this case, there must be a bomb blocking B, but our photon passed by C. In other words, we have ‘seen’ the bomb without a photon ever touching it.

J.P.D.

with photons' or 'not imaging the hair at all'^{4,5}.

One might think that the same Zeno trick could be used to improve counterfactual quantum computers, so that they could actually be useful for something — or at least, that the probability of their finding the right answer should be better than that achieved by flipping coins or tossing dice. But this hope proved misguided: the straightforward application of the Zeno idea to counterfactual quantum computing did not seem to beat the random-guessing limit, and things were at a standstill.

Enter Hosten *et al.*¹. These authors have discovered that chaining Zeno boosters together produces a super Zeno booster that is indeed capable of beating the random-guessing limit in counterfactual quantum computing. They also demonstrate that the idea works with a quantum-optical implementation of the Grover search algorithm. This algorithm proves that a quantum computer can always perform better than any classical computer in searching an unsorted database for a target datum⁷. It has a particularly simple implementation in

quantum optical interferometry, as does the super-sized quantum Zeno effect. Hosten and colleagues show how the two can be married to locate the target element in a database — without running the quantum computer that searches for that element in the first place.

So what is counterfactual quantum computing good for, other than to illustrate the conclusion of a rather obscure quantum paradox? The computer must still be programmed and turned on, even if it is not run, so the approach will not save on electricity bills or labour costs. In fact, the value of the experiment lies simply in furthering our understanding of quantum mechanics and its interface with computation. The entire field of quantum information arose from physicists trying to understand the implications of paradoxes in quantum theory, and already the field is beginning to evolve on its own. The first commercial applications of quantum information, such as real-world quantum cryptography, are now being deployed.

A century ago, we saw the start of the first quantum revolution — the discovery of the

quantum rules that underpin our world. We are now on the verge of a second revolution⁸, in which these rules spawn technological applications. Results such as those of Hosten and colleagues¹ are significant markers on the road to that revolution. ■

Jonathan P. Dowling is at the Hearne Institute for Theoretical Physics, Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803, USA, and the Institute for Quantum Studies, Texas A&M University.
e-mail: jdowling@lsu.edu

1. Hosten, O., Rakher, M. T., Barreiro, J. T., Peters, N. A. & Kwiat, P. G. *Nature* **439**, 949–952 (2006).
2. Elitzur, A. C. & Vaidman, L. *Found. Phys.* **23**, 987–997 (1993).
3. Mitchison, G. & Jozsa, R. *Proc. R. Soc. Lond. A* **457**, 1175–1193 (2001).
4. Kwiat, P. *et al.* *Phys. Rev. Lett.* **74**, 4763–4766 (1995).
5. Kwiat, P., Weinfurter, H. & Zeilinger, A. *Sci. Am.* **275**(5), 72–78 (1996).
6. Knight, P. *Nature* **344**, 493–494 (1990).
7. Grover, L. K. *Phys. Rev. Lett.* **79**, 325–328 (1997).
8. Dowling, J. P. & Milburn, G. J. *Phil. Trans. R. Soc. Lond. A* **361**, 1655–1674 (2003).

NEUROBIOLOGY

Efficiency measures

Michael R. DeWeese and Anthony Zador

The nervous system translates sensory information into electrical impulses. The neural 'code' involved seems to represent natural sounds and images efficiently, using the smallest number of impulses.

Our perception of the outside world relies on the transformation of physical signals (such as light and sound) into a pattern of neural impulses, or spikes. These spikes are then transmitted to higher brain regions, where they are further transformed into other patterns of sensory spikes, and ultimately into the motor spikes that mediate behaviour. What is the relationship (the 'neural code') between these neural responses and the sensory signals they represent? Are there general principles underlying the neural code?

The 'efficient-coding hypothesis'¹ proposes that sensory neurons are adapted to the statistical properties of sensory signals to which neurons are exposed. Two papers in this issue invoke this principle to predict how neurons encode natural auditory and visual stimuli, as opposed to the artificial stimuli often used in experiments. Smith and Lewicki (page 978)² develop an algorithm to find an efficient representation of natural sounds and speech, and show that this theoretically predicted representation matches that observed experimentally in the auditory nerve of cats. Sharpee and colleagues (page 936)³ show that cortical neurons adapt over seconds or minutes during the course of an experiment to maximize the information they provide about the stimulus.

Together, the two papers show how the efficient-coding hypothesis can help to make sense of properties of the neural code on both evolutionary and behavioural timescales.

In the cochlea, sound is encoded into spikes, which are transmitted along the auditory nerve to higher stations in the auditory system. Auditory nerve fibres each respond to a narrow range of sound frequencies, with the range generally increasing with the median frequency. The response of each auditory nerve fibre can therefore be modelled as a (nonlinear) 'filter' that removes frequencies outside a particular range. Why do the auditory nerve filters have the particular form they do? Smith and Lewicki reasoned that if the auditory code is indeed 'efficient', then they should be able to predict the form of the auditory filter bank by finding the sparsest code; that is, the one that requires the least activity.

To obtain this prediction, Smith and Lewicki first expressed the efficient-coding hypothesis as an algorithm whose input is an ensemble of sounds, and whose output is a sparse encoding for transmitting or representing this ensemble. The algorithm discovers that the sparsest encoding of sounds is into brief events suggestive of spikes, the precise timing of which conveys much of the information. The sparsest

code depends on the ensemble of sounds to be encoded; a code that is most efficient for one set of sounds is not necessarily most efficient for another.

Why should the most efficient code depend on the stimulus ensemble? The basic intuition is straightforward. Suppose I ask you to describe individual sounds produced by different musical instruments, but I limit your vocabulary to only four words (of your choosing). If you know that the instruments are used in a rock band (that is, they are chosen from the rock ensemble), you might choose a code consisting of the words 'guitar', 'bass', 'drums', 'key-board'; but if the instruments are used in a classical orchestra (the classical ensemble), you might choose instead 'woodwind', 'brass', 'percussion', 'string'. So, the choice of the most efficient code depends on what is being described.

The dependence of the sparsest code on the ensemble raises the question of exactly which ensemble should be considered the evolutionarily relevant 'natural' ensemble. Smith and Lewicki therefore tested three subsets of sounds: animal vocalizations, transient environmental sounds such as crunching leaves and cracking twigs, and ambient environmental sounds like rain. They found that if the ensemble consisted of any single sub-ensemble, then the predicted sparse code failed to match previous experimental observations of cat auditory filters. However, the code predicted from a carefully balanced mixture of these natural sounds provided a good match to the experimental data, supporting the idea that these filters evolved to encode natural sounds efficiently. Interestingly, filters calculated from a human speech ensemble also fit the cat data well; because (as any cat owner will attest) the cat auditory system did not evolve to

attend to human speech, the authors speculate that speech may have evolved to match the properties of cochlear filters.

The efficient-coding hypothesis posits that the neural code should be adapted to the statistical properties of stimuli on all timescales. Where Smith and Lewicki were studying adaptation that had occurred over an evolutionary timescale, Sharpee and colleagues³ looked for adaptation over seconds and minutes, the much faster timescale relevant to behaviour.

Sharpee and colleagues studied adaptation to natural scenes (images of a forest) of a particular subclass of neurons — the 'simple cells' — in the primary visual cortex of anaesthetized cats. Simple cells respond to oriented bars or edges, and it was already known that they could adapt to the most basic statistical properties of images — the luminance, mean and contrast (variance). So the authors probed these neurons with a control ensemble consisting of white-noise images (like the static on an untuned television monitor) whose mean and contrast were matched to those of the natural scenes. Consistent with the efficient-coding hypothesis they found that, after seconds or minutes of exposure to the natural ensemble, these neurons adapted their response properties so as to increase the information they transmitted.

The appeal of the efficient-coding hypothesis is that it predicts structure in a large and complex data set: the response properties of sensory neurons throughout the nervous system. As such, it is one of the great successes of theoretical neuroscience. But why should the nervous system encode sensory stimuli efficiently? One possible explanation is that neuronal connectivity requires space, and that therefore information must be transmitted using as few 'wires' (axons) as possible⁴. Alternatively, the limiting resource may be the energy associated with neuronal activity⁵; a sparser code using fewer spikes uses less energy.

Finally, the motivation might (as Barlow¹ supposed) be computational: sparse encoding requires uncovering as much of the underlying structure of the signals as possible. That the nervous system seems to be able to do this is impressive, because it is not an easy task. As anyone who has ever tried to write a *Nature* letter knows, writing succinctly requires a very clear idea of what one is trying to say. Or as Mark Twain put it, "I'm sorry this letter is so long, but I did not have time to make it shorter."

Michael R. DeWeese and Anthony Zador are in the Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. e-mails: deweese@phage.cshl.org; zador@cschl.org

1. Barlow, H. B. in *Information Processing in the Nervous System* (ed. Leibovic, K. N.) 209–230 (Springer, New York, 1969).
2. Smith, E. C. & Lewicki, M. S. *Nature* **439**, 978–982 (2006).
3. Sharpee, T. O. *et al.* *Nature* **439**, 936–942 (2006).
4. Chklovskii, D. B., Schikorski, T. & Stevens, C. F. *Neuron* **34**, 341–347 (2002).
5. Laughlin, S. B., de Ruyter van Steveninck, R. R. & Anderson, J. C. *Nature Neurosci.* **1**, 36–41 (1998).

PARTICLE PHYSICS

Quarks on a gravitational string

Nick Evans

Quantum chromodynamics, the theory of the strong nuclear force, is notoriously intractable. An alternative approach brings gravity to bear, and produces fairly accurate predictions of some physical quantities.

The strong nuclear force is the force that causes quarks to bind together to form composite particles, such as the proton. It is explained within the standard model of particle physics by a theory known as quantum chromodynamics (QCD) in terms of fields analogous to electric fields that arise between particles that possess 'colour' charge — the strong-force equivalent of electric charge. Unfortunately for the theorists, however, QCD has consistently eluded analytical solution. The best available calculations rely on huge supercomputer simulations, and the parameters that emerge must be fitted to experimental results.

That situation might change with a remarkable complementary description of the strong force. The new theory incorporates four spatial directions and a dynamic force of gravity that curves space-time in the spirit of the general theory of relativity. As Erlich *et al.*¹ writing in *Physical Review Letters* and Da Rold and Pomarol² writing in *Nuclear Physics B* report, it can already be used to produce reasonably accurate quantitative descriptions of light quark–antiquark pairings (mesons) such as the pion.

The genesis of the models is string theory, in which fundamental particles are not points in space, but have an intrinsic length. Seen from afar, the length of these one-dimensional 'strings' is too small to be discerned, but oscillations along this length determine the mass and spin of the particle the string represents. String theory has become a leading contender for a theory of everything, as it can accommodate both gravity and the other forces of nature: so-called open strings have free ends and look like the photon or gluon (the particles that mediate the electromagnetic and strong nuclear force as quanta of the respective fields), whereas closed loops of string have the properties of a graviton, the as yet undiscovered particle that would mediate gravity in a quantum version of general relativity. Problematically, however, string theory turns out to make sense only in a world with nine spatial dimensions, rather than the three that we see.

String theory also contains objects of dimension larger than one: two-dimensional 'membranes', and extensions of still higher dimension known as branes (Fig. 1). These branes emerge in the theory as sub-spaces to which the motion of the open strings, those relevant to QCD, is restricted. Much entertainment has been derived from constructing theories with different matter fields and forces

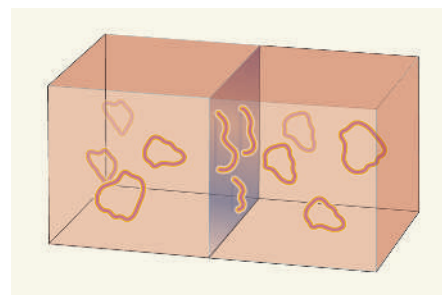


Figure 1 | Braney stuff. In the string theory construction of quantum chromodynamics (QCD), open strings, which describe quarks, are restricted to a membrane-like surface, known as a brane. Closed strings live in a higher-dimension space off the brane and generate an alternative description of QCD.

restricted to the volumes of branes of different dimensions — including three.

In brane theories, the closed strings that describe gravity continue to move in all nine spatial dimensions. Away from the branes' surfaces, the branes' energy generates curvature in space-time, as does gravity in general relativity. Naively, there are therefore two sectors to brane theories: off the brane, describing a theory with similarities to gravity, and on it, describing a theory such as QCD.

But what if these two sectors are actually two alternative descriptions of the same physics? This was originally conjectured³ for an idealized version of QCD incorporating constraints on its solution to make the theory easier to handle mathematically. But subsequent work that has aimed to remove these devices seems also to support the supposition.

To appreciate how two different theories can describe the same physics, we must first understand that the behaviour of QCD is very different at different energy scales. Colour charge depends, for instance, on the energy exchanged between two quarks in an interaction. If the energy is large, the coupling is weak; if it is small, the coupling becomes very strong. So, the farther quarks move away from each other (and the less energy is exchanged between them), the stronger they bind. This 'quark confinement' is the reason why we never see a free quark, only those bound into composite particles.

In the gravitational, string-theory description of this phenomenon, the quark coupling is represented by a field that describes the distribution of strings in space. The role of energy scale is assumed by the direction radial from

the brane in the extra dimensions of the purely gravitational theory. Variations in quark coupling turn up as a variation in a gravity field in this radial direction. Crucially, gravity fields are weakly coupled and therefore easy to compute — unlike the fields of QCD.

Another consequence of the strong coupling in QCD is that its vacuum is actually filled with quark–antiquark pairs, created by ‘borrowing’ energy for a short period of time, as permitted by the Heisenberg uncertainty principle. The quark and antiquark might be expected to annihilate almost instantaneously, but the strong force is so strong that the energy liberated by their attraction is greater than that borrowed by the vacuum to create them. The vacuum’s debt can thus be paid off. In the gravitational string-theory approach, the vacuum quark density is, like quark coupling, described by a field in the higher-dimension space. The solution for this field switches on at values of the radius where the coupling is strong⁴ — an important first step in showing that the approach indeed can reproduce the same behaviour as does QCD.

Quark–antiquark pairings such as pions are simply fluctuations in the number of quarks above vacuum level. In the gravitational description, the known mesons are wave excitations on top of the background field configuration just described. However, the mathematical tricks used to get to this stage mean that there are more stable wave excitations than there are mesons predicted by QCD, corresponding to there being extra heavy quarks present. Erlich *et al.*¹ and Da Rold and Pomarol² study the wave excitations appropriate to known QCD mesons. (Only particular waves thrown up by the gravitational theory, corresponding to discrete meson mass values, are stable in space-time.)

The gravitational theory has the same number of defining parameters as QCD. Fixing the parameter that determines the warping of space-time in the gravitational theory corresponds to fixing the QCD coupling to its experimentally measured value. The authors use the background value of a gravitational field to set the masses of the two lightest quarks, ‘up’ and ‘down’, through the mass of the various pions and the rho meson (each of which is a combination of an up and down quark and an up or down antiquark). They can use these masses to predict the masses of other bound particles and the strengths of interactions between these states. The values they come up with lie within 10–15% of the values found in nature.

The agreement suggests not only that the approach could provide a radical new description of QCD phenomena such as the mass spectrum of the particles found, but also that many of the consequences of strong-force interactions are common to a broad range of theories. The outstanding challenge is to find a way to remove all artefacts of the mathematics used to simplify string theory in a controlled fashion.

There is indeed one arena of QCD where

the gravitational description provides our current best theoretical tool. In collisions between heavy nuclei, the temperature and density are sufficient to compress nuclei into a soup of quarks. Properties of this strongly interacting fluid, such as its viscosity and thermal conductivity, are very hard to compute in QCD, but can be extracted with relative ease from the gravitational theory⁵.

It seems we have a new tool in our kit for tackling the strong force and the complexities of quarks. ■

Nick Evans is in the School of Physics and Astronomy, University of Southampton, Highfield, Southampton SO17 1BJ, UK. e-mail: evans@phys.soton.ac.uk

1. Erlich, J., Katz, E., Son, D. T. & Stephanov, M. A. *Phys. Rev. Lett.* **95**, 261602 (2005).
2. Da Rold, R. & Pomarol, A. *Nucl. Phys. B* **721**, 79–97 (2005).
3. Maldacena, J. M. *Adv. Theor. Math. Phys.* **2**, 231–252 (1998).
4. Babington, J., Erdmenger, J., Evans, N. J., Guralnik, Z. & Kirsch, I. *Phys. Rev. D* **69**, 066007 (2004).
5. Son, D. T. & Starinets, A. O. preprint available at <http://arxiv.org/abs/hep-th/0601157> (2006).

EVOLUTION

Careful with that amphioxus

Henry Gee

The textbook tale of vertebrate origins is brought into question by phylogenetic analyses of new genomic data. But the amphioxus, long viewed as a precursor to fish, remains a central character in events.

History is written by the victors. This is as true for our account of evolution as it is for purely human affairs. But as the paper by Delsuc *et al.* (page 965 of this issue¹) makes plain, we the apparent victors still need to be prepared to rethink our own deep evolutionary history as time and technology advance.

The conventional picture of the evolution of the deuterostomes (our particular corner of organized life) has been one of steady and anthropocentric advancement (Fig. 1a). Starting with undistinguished squashy sea creatures, evolution produced the first signs of gill slits (in hemichordates, or acorn worms); a dorsal tubular nerve cord and a notochord (tunicates, or sea-squirts); and clear muscular segmentation (cephalochordates, or lancelets, represented by the fish-like amphioxus). The culmination was vertebrates (with all of the above characteristics, and heads), and finally ourselves. The echinoderms (starfishes, sea-urchins and so on), with their peculiar symmetry and strange calcitic skeletons, are seen as bizarre relatives to be locked in the attic,

rather like the first Mrs Rochester in Charlotte Brontë’s *Jane Eyre*.

Time and again, further work has exposed our prejudices for the parochial conceits that they are. As long ago as 1881, it was proposed that hemichordates and echinoderms formed a discrete group, the Ambulacraria², a proposal revived at first tentatively and then supported with increasing conviction by molecular evidence (Fig. 1b)³. And there have been persistent signs, from fragments of both morphological and molecular evidence, that the similarities between amphioxus and vertebrates conceal a wealth of difference — while the manifest oddities of tunicate morphology and development might be more attributable to recent innovation than to ancient heritage⁴.

In their report, Delsuc *et al.*¹ apply a range of phylogenetic methods to a large genomic data set from an unrivalled range of taxa. They control for known problems such as ‘long-branch attraction’, and show that tunicates — not lancelets — are the closest extant relatives of

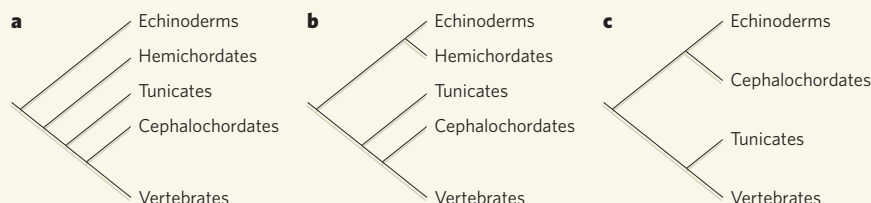


Figure 1 | Deuterostome relationships. **a**, The classic, textbook view, implying a smooth increase in complexity from a relatively simple and sedentary deuterostome ancestor to motile vertebrates. **b**, A more recent view informed by molecular evidence, in which hemichordates are allied with echinoderms, implying a more complex echinoderm history. **c**, The topology suggested by the results of Delsuc *et al.*¹. This implies that the deuterostome ancestor would have been motile and relatively complex, and that the sessile habits of most echinoderms and tunicates evolved later. Hemichordates are notably absent.

vertebrates (Fig. 1c). Moreover, they argue that amphioxus is not a close relative of either vertebrates or tunicates, but is more akin to echinoderms. This conclusion is not nearly as robust as the tunicate–vertebrate link. It does, however, seem to be a persistent feature of the analysis, and is worth careful consideration. After all, the close hemichordate–echinoderm link seemed unorthodox when it was first revived, and few would have expected that the tunicate–vertebrate link would receive such strong support.

So, if lancelets really are close relatives of echinoderms, what are the implications for our picture of deuterostome evolution? The short answer is that the textbook scheme is turned on its head. Rather than the steady acquisition of progressively more chordate-like (and, by implication, human-like) features from an ancestor with nothing much to recommend it, the story becomes one of persistent loss. The last common ancestor of extant deuterostomes would have been a free-living, bilaterally symmetrical creature with a distinct throat region perforated by gill slits, segmented body-wall musculature and possibly a reasonably sophisticated brain and central nervous system. In a sentence, the ancestor would have looked like a cross between an amphioxus and a larger, brainier, tunicate tadpole larva. Crazy? Possibly. But possibly not.

Most modern tunicates seem to be immobile relatives of motile ancestors, but some, such as the appendicularian tunicate *Oikopleura*, do move about. So in this group the proposed ancestral deuterostome portrait has not so much been erased as extensively modified by changes in the evolution of tunicate development. Fossil evidence fills in more details, with the discovery of the strange Cambrian vetulicolians⁵, which look strikingly like the hypothetical ancestral deuterostome form — and, like the 'somatico-visceral animal', a model for a chordate ancestor described in one of the last and most prescient papers by the late great palaeontologist A. S. Romer⁶.

Echinoderms have lost all these features to the extent that the ancestral portrait cannot be recognized at all in modern forms. However, some members of a group of fossil echinoderms called stylophora may have had gill slits (an interpretation made more likely by the re-affirmed hemichordate–echinoderm link) and, according to R. P. S. Jefferies' controversial calcichordate theory, an internal organization similar to that of tunicates⁷. Jefferies proposes that each kind of stylophoran should rather be assigned to the groups containing modern tunicates, vertebrates or cephalochordates, respectively, and that the characteristic echinoderm calcite skeleton has been lost in each case. However, a simpler alternative is that the calcite skeleton was acquired just once in the ancestry of echinoderms, when this creature still looked much like the proposed deuterostome ancestor. The irony is that even with this shift in perspective, many of Jefferies' detailed anatomical

interpretations of stylophoran anatomy, particularly as regards the anatomy of the pharynx⁸, could still be correct (but see ref. 9).

Much remains to be found. As Delsuc *et al.*¹ acknowledge, adequate tests of these ideas require the sequenced genomes of hemichordates, of more echinoderms and — especially — of amphioxus. This last is particularly important: for generations, amphioxus has been viewed as a model organism, representing a picture of the first stirrings of vertebrate evolution. But if this radical and certainly controversial new view is supported by further evidence, amphioxus could occupy a far more significant and inclusive position — the

closest extant organism we yet have to the ancestor of all deuterostomes. ■

Henry Gee is a senior editor at *Nature*.

1. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. *Nature* **439**, 965–968 (2006).
2. Metchnikoff, V. E. *Zool. Anz.* **4**, 139–157 (1881).
3. Halanych, K. M. *Mol. Phylogenet. Evol.* **4**, 72–76 (1995).
4. Gee, H. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 1–14 (Taylor & Francis, London, 2001).
5. Shu, D.-G. *et al. Nature* **414**, 419–424 (2001).
6. Romer, A. S. *Evol. Biol.* **6**, 121–156 (1972).
7. Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, London, 1996).
8. Dominguez, P., Jacobson, A. G. & Jefferies, R. P. S. *Nature* **417**, 841–844 (2002).
9. Clausen, S. & Smith, A. B. *Nature* **438**, 351–354 (2005).

PLANETARY SCIENCE

Pluto's expanding brood

Richard P. Binzel

Pluto is no lone ranger in the farthest expanses of the Solar System — its travelling companions now number three. And if Pluto can have so many, why shouldn't other objects in the distant, icy Kuiper belt?

Once thought to be a solitary denizen of the outer reaches of the Solar System, Pluto — which piqued our curiosity in 1978 with the discovery of its large satellite, Charon¹ — is becoming ever more intriguing. In fact, the relative sizes of Pluto and Charon (Charon's diameter of around 1,200 kilometres is just over half that of Pluto's) means they are a 'double planet', orbiting a mutual centre of gravity, or barycentre, outside the surface of Pluto. But the story does not stop there. On page 943 of this issue, Weaver *et al.*² present Hubble Space Telescope images showing that the Pluto system is at least quadruple. And as Stern *et al.* indicate in a companion paper³ on page 946,

this complexity portends further discoveries: more small satellites may be lurking out there, and cratering impacts on them may have liberated rings or arcs of matter. Propitiously, NASA's New Horizons mission^{4,5} is now successfully launched (Fig. 1) and on its way to a flying visit to Pluto and its companions in 2015.

Following Clyde Tombaugh's discovery of Pluto in 1930, searching for satellites was an obvious first task. But none was found until Pluto's march towards its point of closest approach to the Sun, coupled with the exquisite optics of a ground-based telescope, finally allowed Charon to be pinpointed¹. Since then, ground-based surveys^{6,7} have yielded



Figure 1 | Destination Pluto. The New Horizons spacecraft took off from Cape Canaveral on 19 January 2006 aboard an Atlas V rocket, bound for the Pluto system. Speedy results are not to be expected: the half-tonne, piano-sized spacecraft must cover a distance of just under five billion kilometres, and will reach a point of closest approach some 10,000 kilometres from Pluto on 14 July 2015.

NASA/ESA/HST

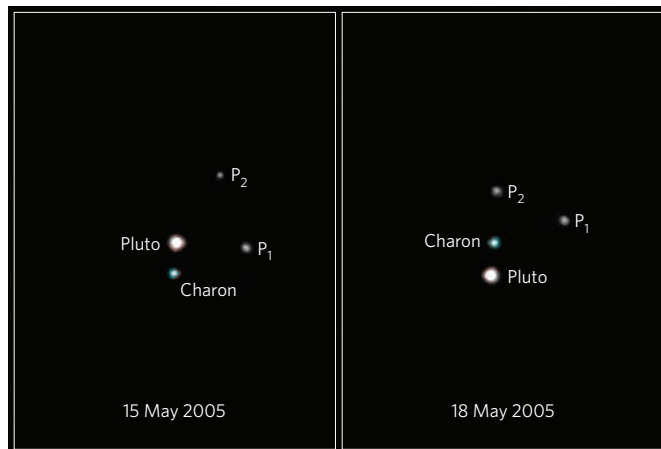


Figure 2 | A remote quartet. Two Hubble Space Telescope images of the Pluto system taken three days apart, revealing the existence of two smaller satellites, P1 and P2, in addition to Charon (discovered in 1978). P1 is the farther of the two newcomers from the system's centre of gravity, which lies just above the surface of Pluto. It completes just one orbit for every six of Charon's; P2 completes about one-and-a-half in the same time.

no evidence for other satellites larger than about 160 kilometres in diameter. Motivated by the impending launch of New Horizons, Weaver and colleagues² secured Hubble Space Telescope time in May 2005 to search for satellites as small as around 25 kilometres across⁸. Sure enough, two objects were found travelling through space with Pluto that had motions consistent with orbits around the Pluto–Charon barycentre (Fig. 2).

This discovery² prompted a fresh look at Hubble images from 2002 taken to map Pluto's surface. Although these images were not optimized for the identification of satellites, when preliminary orbital calculations gleaned from the 2005 images were added in, the presence of two additional companions was confirmed. The diameters of the satellites, creatively dubbed 'P1' and 'P2' (they will receive their official names later this year), are respectively around 60 and 50 kilometres, assuming surface reflectivities similar to that of Charon. (They are larger if their reflectivities are lower.)

Satisfying as discovery for discovery's sake is, it is the intriguing orbits of the newly spotted satellites that is creating the most scientific excitement. The present, limited data show that P1 and P2 are in circular orbits in the same plane as Charon. Moreover, the radii of their orbits place them in a resonant dance with Charon: for every twelve orbits Charon makes, P1 completes almost exactly two; in the same time, P2 (which is closer in) completes nearly three. Such consonance is not likely if P1 and P2 are captured objects that just happened, once upon a time, to have ventured too close to Pluto: tidal forces from Pluto and Charon are not great enough to coerce captured objects into co-planar resonances over the age of the Solar System³. The most plausible explanation is that Charon, P1 and P2 are all Pluto's progeny, and split off from it through a giant impact^{3,9}. The disk of material ejected by this collision into orbit around Pluto allowed these satellites — and perhaps others yet unseen — to condense in co-planar, circular orbits³. The resonant niches occupied by P1 and P2 may have been particularly fertile locations for coalescing material, or for maintaining long-term orbital stability.

As Stern and colleagues point out³, implications abound for Pluto's brethren in the Kuiper belt, the disk-shaped region of small, icy bodies found outside the orbit of Neptune. Within current detection limits, up to a fifth of all Kuiper-belt objects seem to have satellites or to be part of a binary system¹⁰. Pluto is the first known quadruple system, but multiple companions may be just the tip of the iceberg for the complexities of gravity's play on small bodies far from the perturbative forces of the Sun and giant planets. For example, most ejected debris from cratering impacts on P1 and P2 can easily escape the satellites' surfaces, but not the gravitational hold of the Pluto system. So tenuous rings or ring arcs may be the rule, rather than the exception, for Pluto and other multiple-bodied congregations in the Kuiper belt³. Even quadruple systems may become *passé* as investigations become increasingly percipient.

Those planning NASA's New Horizons mission, now en route first to a gravity assist from Jupiter in February 2007 and then its July 2015 appointment with Pluto, are now adding to their to-do list highly resolved imaging and spectroscopy of the newly discovered satellites. Refining these satellites' sizes and their orbital positions in nine years' time will also be a priority for observations to follow those currently being reported². Both on its way in and out of the Pluto system, New Horizons' instruments will canvass the orbit plane for more satellites, rings and other telltale signs that might reveal the origin and evolution of this close-knit family. Pluto is a lonely place no more. ■

Richard P. Binzel is in the Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mail: rpb@mit.edu

- Christy, J. W. & Harrington, R. S. *Astron. J.* **83**, 1005–1008 (1978).
- Weaver, H. A. et al. *Nature* **439**, 943–945 (2006).
- Stern, S. A. et al. *Nature* **439**, 946–948 (2006).
- <http://pluto.jhuapl.edu>
- www.nature.com/news/2006/060116/full/060116-2.html (2006).
- Stern, S. A. et al. *Icarus* **94**, 246–249 (1991).
- Stern, S. A. et al. *Icarus* **108**, 234–242 (1994).
- Steffl, A. J. et al. *Astron. J.* (submitted); preprint available at <http://arxiv.org/abs/astro-ph/0511837> (2005).
- Canup, R. M. *Science* **307**, 546–550 (2005).
- Stephens, D. C. & Knoll, K. S. *Astron. J.* **131**, 1142–1148 (2006).



50 YEARS AGO

In his Stephen Paget Memorial Lecture, "Humanity's Rising Debt to Medical Research", Sir Henry Dale referred to the convincing evidence of the success of inoculation against diphtheria... Systematic inoculation began in Hamilton, Ontario, in 1925, and the death-rate from diphtheria fell so steeply that after 1930 there were no more deaths and after 1933 no more cases. In Great Britain a full-scale preventive inoculation campaign was not inaugurated by the Ministry of Health until 1940 and was not fully under way until 1942. In 1940 there were still in England more than 45,000 cases of diphtheria, and more than 2,400 of these were fatal. As inoculation became effective, in spite of a counter-campaign, the numbers fell steeply and steadily until in 1954 there were only 173 cases; of the nine deaths, six were in children less than fifteen years old, and all in the minority who had not been inoculated.

From *Nature* 25 February 1956.

100 YEARS AGO

The traditional scientific man has disappeared almost as completely as the traditional Yankee of the stage. The change came gradually, but the proof that it had come was brought before us suddenly. In 1902 there was called in New York a meeting of those who were designated by the picturesque expression captains of industry. To that meeting representatives of science were invited, not as lions to be stared at, but to sit with the leaders of the industrial and commercial world as representatives of science, and not only of applied science, but of pure science. As the captains of industry were supposed to be men of force in organising and to have a keen insight into men and things, we had a right to feel that science had been honoured, perhaps not more than ever before, but for a reason for which it had not been honoured before in the United States... The conception of a scientific man as a captain of industry means simply the acknowledgement that science has a practical relation to the world.

From *Nature* 22 February 1906.

50 & 100 YEARS AGO

MARINE BIOLOGY

Spawning spot

The enigma of the eel life-cycle has become a little less enigmatic — at least as far as the Japanese eel, *Anguilla japonica*, is concerned. As described elsewhere in this issue (*Nature* **439**, 929; 2006), Katsumi Tsukamoto has filled in a missing part of the picture of the species' ocean biology.

When they are mature, eels migrate from fresh water out into the open ocean, travelling thousands of kilometres to spawn. But the when and where of the actual spawning event, although suspected in general terms, have been hard to pin down. This is partly because of the technical challenges of isolating and identifying the tiny, transparent larvae just after they hatch. Tsukamoto has tackled those



challenges to pinpoint the timing — which aptly coincided with a new moon in June — and strategic location of *A. japonica* spawning.

This fishing feat also meant that the newly hatched larvae could be photographed. As shown in this picture, teeth, jaws and pigmented eyes are evident within five days of hatching, when the larvae are still less than 5 millimetres long.

Thanks to the efforts in the 1920s of a Danish oceanographer, Johannes Schmidt, the spawning

area of the European (*A. anguilla*) and American (*A. rostrata*) eels was identified as the Sargasso Sea in the western North Atlantic, spreading over some 5 million square kilometres. But it seems that the nursery of the Japanese eel, as identified by Tsukamoto, is much more compact — spawning is confined to a seamount northwest of Guam in the Philippine Sea, a location that enables the larvae to ride the northbound Kuroshio Current

to their freshwater destinations in eastern Asia.

This discovery will be applauded by researchers specializing in eel biology, as well as by those intrigued by a life history that constitutes one of nature's wonders. But there is a commercial aspect too: eels are of great economic value, not least in Asia, and there is a pressing need for management and conservation of natural populations.

Rosalind Cotter

OCEAN RES. INST., UNIV. TOKYO

PARASITOLOGY

Switching like for like

Piet Borst and Paul-André Genest

To remain hidden from its host's immune system, the malaria parasite must vary the proteins on the surface of the infected cell. The genes encoding these proteins are very similar, so how does the parasite express just one at a time?

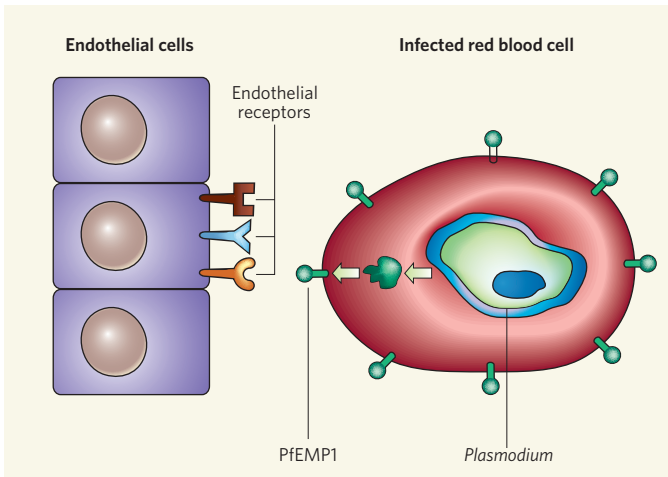
Selecting one gene for activation out of many closely similar ones is tricky. Mammals address this problem when they switch on just one odorant receptor gene in each olfactory neuron, choosing between a thousand receptor genes. Bacteria and protozoa address it when they vary the proteins displayed on their surface. How an organism can activate just one gene and keep all the other closely related ones silent — a process known as allelic exclusion — has long remained elusive. In this issue, Voss *et al.* (page 1004)¹ analyse how this selection process works for the *var* gene family of the malaria parasite *Plasmodium falciparum*.

The *var* genes are expressed when the parasite multiplies in the red blood cells (erythrocytes) of its host. The genes encode proteins known as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), so-called because they are exported from the parasite and end up in the membrane of the red blood cell (Fig. 1). PfEMP1 is key to the lethality of *P. falciparum* infection, because the PfEMP1 molecules on the erythrocyte surface cause the cell to stick to capillary walls. In this way, the infected cells are prevented from circulating through the spleen where they would be recognized by immune cells and removed from the circulation. So PfEMP1 provides the parasite with more time

to multiply. The unfortunate side effect is that capillaries become clogged, resulting in the severe symptoms of tropical malaria.

Putting 'alien' proteins, or antigens, on the erythrocyte surface inevitably advertises the parasite's presence to the host immune system, and eventually results in an immune response against the infected cell. To escape this threat, the parasite uses antigenic variation of its PfEMP1. The *var* family consists of 60 genes, each encoding a different PfEMP1.

Figure 1 | A red blood cell infected with *Plasmodium falciparum*. The surface of the infected cell is studded with proteins of the PfEMP1 family, encoded by *var* genes. The various PfEMP1 molecules can bind to one of a series of receptors on the endothelial cells that line the capillary walls. (Modified from ref. 14.)



By switching the *var* gene expressed, the parasite changes the PfEMP1 on the host-cell surface before the host immune system can mount a full response to it. This allows the parasite to survive and multiply for another week.

Antigenic variation of surface PfEMP1 works in practice only if the parasite expresses just one *var* gene at a time. If all 60 genes were simultaneously expressed, the host would make antibodies against all 60 at once and eliminate the infected cells. The parasite must also be able to switch the *var* gene expressed, otherwise it would continue to make the same surface antigen and be doomed by the host immune response. Key elements in the strategy of antigenic variation are therefore a large gene family, the ability to activate only one gene and keep the others silent (allelic exclusion), and the ability to switch the gene that is active².

The *var* genes encode large proteins (relative molecular mass 250,000–300,000) and have

two regulatory elements that are known to affect gene silencing: a 'promoter', which lies before the start of the protein-coding portion of the gene, and a large intron (a gene segment that is transcribed into RNA, but is deleted from the transcript before the final RNA is translated into protein). To study the expression of *var* genes, Voss *et al.*¹ used an artificial marker gene as a proxy for the *var* gene. The expression of this gene, called *dhfr* (for dihydrofolate reductase), was put under the control of a *var* gene promoter (*upsC*). Parasites in which the *dhfr* marker gene was activated were selected by culturing the cells with an antifolate drug. Parasites in which the gene was active were resistant to the drug, whereas parasites with a silent *dhfr* gene died.

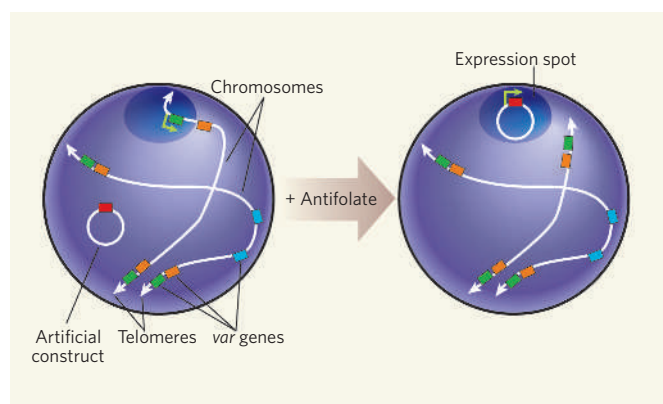
Although the artificial gene was present on a plasmid — a small extra circle of DNA — it behaved like the parasite's own (endogenous) *var* genes, occasionally switching on and off spontaneously. The most interesting result, however, was that the activation of the *upsC*-*dhfr* gene-construct silenced the previously active endogenous *var* genes. The artificial gene seemed to become an authentic member of the *var* gene family, able to compete with the other members to be the actively transcribed gene. Thus, the active *upsC* promoter seems to be sufficient to silence expression by competing *var* genes. Dzikowski *et al.*³ have obtained similar results with different artificial *var* gene-constructs.

There is still controversy about the role of the enigmatic *var* gene intron in silencing. This intron is present in all *var* genes and contains a promoter yielding large amounts of RNA transcripts that are not translated into protein. Previous reports concluded that the intron is essential for allelic exclusion to occur^{4–6}. But Voss *et al.*¹ found that the absence of the intron did not affect either the ability to switch genes or the ability to keep the inactivated gene completely silent. The intron only decreased the rate at which the silent gene was activated and also seemed to limit the activity of neighbouring genes (a 'boundary effect'). However, Voss and colleagues' artificial gene-construct¹ also contains a second promoter; this drives the expression of a blasticidin resistance gene that was used to introduce the *upsC*-*dhfr* construct into the *Plasmodium* cells. It is therefore possible that this second promoter can replace the intron promoter to allow complete silencing of the *upsC* promoter. That would explain why Voss *et al.*¹ could dispense with the intron, whereas other groups^{3–6} could not.

Voss *et al.* argue that regulation of *var* gene expression is relatively simple (Fig. 2). In their model, the *upsC* promoter recruits proteins that keep the *var* gene silent⁷. The gene can be activated by moving to a unique location in the nucleus⁸. This 'expression spot' can accommodate only a single active gene promoter, explaining how all the other *var* gene promoters are kept silent.

Figure 2 | A model for the regulation of *var* genes in the *Plasmodium* nucleus.

Voss *et al.*¹ propose that a *var* gene (here represented by the artificial gene in a plasmid) is activated by moving to an 'expression spot' where it displaces the previously active endogenous *var* gene. Both active and inactive *var* genes are found at the nuclear periphery⁸. (Modified from ref. 1.)



Voss and colleagues' model for regulating *var* gene expression is similar to that proposed for the control of another parasite gene family, the *vsg* genes that encode the variant surface glycoproteins of African trypanosomes, the parasites responsible for sleeping sickness. Trypanosome *vsg* genes are expressed from about 20 expression sites at the ends of chromosomes (telomeres), only one of which is fully active at a time. When the trypanosome is forced to express two sites (by selection for drug resistance markers), the stressed organism can survive only by rapidly switching between the sites. The two switching expression sites were found to be very close together in the nucleus⁹, indicating the existence of a unique activation area, subsequently characterized¹⁰ as the 'expression site body'.

Although this simple model can explain allelic exclusion in the parasite gene families studied, the control of another well-studied gene family, the odorant receptor genes in mammals, involves an additional step¹¹. The receptor gene family is riddled with 'pseudogenes', and the cell is able to select against expression of these faulty genes. This must mean that there is a feedback in the system: if the activated gene cannot produce a functional gene product, it is switched off and another gene is activated until a functional one is found. Such a feedback is clearly absent in *Plasmodium*, as the artificial gene introduced by Voss *et al.*¹ in the *var* gene family does not encode a functional PfEMP1.

Even for parasites, the expression-site-body model leaves many questions unanswered, as has been stated before¹². What is the nature of the unique site? How do *var* gene promoters get in there, and how does another family member displace them? How are promoters activated in the expression site body? These questions can now be tackled, because the *var* gene family of *Plasmodium* is experimentally more tractable than trypanosomes' *vsg* gene expression sites. Voss and colleagues' experiments¹ used a single marked *var* gene promoter present in multiple copies in the cell. As procedures to introduce genes into *Plasmodium* improve¹³, more precise experiments can be done using two *var* genes, each marked with a different selectable marker, and present in a single copy in the transfected cell. This should

make it possible to determine how allelic exclusion works at the molecular level. ■
Piet Borst and Paul-André Genest are in the Division of Molecular Biology, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. Piet Borst is also in the Department of Medical Biochemistry, Academic Medical Center, University of Amsterdam. e-mail: p.borst@nki.nl

1. Voss, T. S. *et al.* *Nature* **439**, 1004–1008 (2006).
2. Borst, P. in *Antigenic Variation* (eds Craig, A. & Scherf, A.) 1–15 (Elsevier, London, 2003).
3. Dzikowski, R., Frank, M. & Deitsch, K. *PLoS Pathogens* (in the press).
4. Deitsch, K. W., Calderwood, M. S. & Wellem, T. E. *Nature* **412**, 875–876 (2001).
5. Calderwood, M. S., Gannoun-Zaki, L., Wellem, T. E. & Deitsch, K. W. *J. Biol. Chem.* **278**, 34125–34132 (2003).
6. Gannoun-Zaki, L., Jost, A., Mu, J., Deitsch, K. W. & Wellem, T. E. *Eukaryot. Cell* **4**, 490–492 (2005).
7. Freitas-Junior, L. H. *et al.* *Cell* **121**, 25–36 (2005).
8. Ralph, S. A., Scheidig-Benatar, C. & Scherf, A. *Proc. Natl Acad. Sci. USA* **102**, 5414–5419 (2005).
9. Chaves, I., Rudenko, G., Dirks-Mulder, A., Cross, M. & Borst, P. *EMBO J.* **18**, 4846–4855 (1999).
10. Navarro, M. & Gull, K. *Nature* **414**, 759–763 (2001).
11. Shykind, B. M. *Hum. Mol. Genet.* **14** (Suppl. 1), R33–R39 (2005).
12. Borst, P. *Cell* **109**, 5–8 (2002).
13. Janse, C. J. *et al.* *Mol. Biochem. Parasitol.* **145**, 60–70 (2006).
14. Borst, P., Bitter, W., McCulloch, R., Van Leeuwen, F. & Rudenko, G. *Cell* **82**, 1–4 (1995).

CORRECTIONS

The News & Views article "Extinctions: A message from the frogs" by Andrew R. Blaustein and Andy Dobson (*Nature* **439**, 143–144; 2006) stated that the mountain pine beetle *Dendroctonus ponderosae* transmits pine blister rust (*Cronartium ribicola*). In fact, the beetles are not a vector for the rust, but they preferentially attack trees already infected with the fungus.

In "Materials science: Colloids get complex" by Alfons van Blaaderen (*Nature* **439**, 545–546; 2006), parts **d** and **e** of Figure 1 were reversed. The caption and attribution of **d** belong to image **e** in the figure, and vice versa.

In "Physiology: Plants on a different scale" by Lars O. Hedin (*Nature* **439**, 399–400; 2006), reference 15 contained several errors. The correct reference is Yoda, K., Kira, T., Ogawa, H. & Hozumi, K. *J. Biol. Osaka City Univ.* **14**, 107–129 (1963).

OBITUARY

Nicholas Shackleton (1937–2006)

A founding father of palaeoclimatology, and an avid clarinetist.

Much of what we have learned about the dynamics of Earth's climate system has come from the study of ancient climates. In the early 1960s, Nick Shackleton, then a graduate student at the University of Cambridge, UK, developed a mass spectrometer that could analyse the oxygen isotope ratios ($^{18}\text{O}/^{16}\text{O}$) in small numbers of foraminifera, tiny calcareous creatures that can be found fossilized in deep-sea sediments. This innovation triggered a revolution in our understanding of ice-age cycles and provided a cornerstone for the relatively new discipline of palaeoclimatology.

Throughout his career, and until his death on 24 January 2006, Shackleton remained at the forefront of the field, supported by his lab manager Mike Hall. He was elected a Fellow of the Royal Society in 1985 and was knighted in 1998. Among his many awards were the Crafoord prize in 1995 and Japan's Blue Planet prize in 2005. He remained at Cambridge as Professor of Quaternary Palaeoclimatology until his retirement in 2004.

Nicholas John Shackleton studied physics at Clare College, Cambridge, graduating in 1961. In 1967 he received his PhD for a thesis entitled 'The measurement of palaeotemperatures in the Quaternary era'. With his mass spectrometer, Shackleton was able to routinely measure oxygen isotope ratios of both planktonic and the much rarer benthic foraminifera. Given the isotopic co-variation of these organisms in the ocean surface and abyss, Shackleton realized that the dominant control on oxygen-isotope variations was not temperature, as suggested by Cesare Emiliani in the 1950s, but instead was changes in the isotopic composition of the oceans caused by preferential removal of the lighter ^{16}O in the water that makes up continental ice sheets. He had thus discovered a method for reconstructing the history of global (mainly Northern Hemisphere) ice volume through the succession of ice ages.

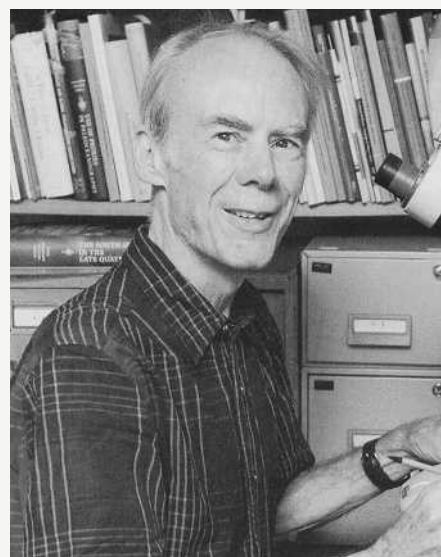
At the end of the 1969 meeting of the International Quaternary Association in Paris, Shackleton found himself the only person in the audience at a talk given by John Imbrie, a palaeontologist. Imbrie presented a statistical method for estimating temperatures using census information on the same species of surface-dwelling foraminifera that Shackleton had been measuring. Shackleton and Imbrie realized that their methods could be combined — Shackleton's isotopic method allowed for the establishment of a temporal

framework based on ice volume, whereas Imbrie's statistical methods would yield the temperature information. Thus, in effect, was born the CLIMAP Project, a multi-institutional programme that in the 1970s produced the first global map of sea surface temperatures during the Last Glacial Maximum, around 21,000 years ago.

In 1973, Shackleton made the fundamental discovery that the dominant ice-volume cycle revealed by the oxygen isotopes lasted roughly 100,000 years. He analysed a sediment core from the western tropical Pacific in which Neil Opdyke had identified the signature of the most recent reversal of Earth's magnetic field, an event that occurred about 780,000 years ago. With an improved timescale that could now be applied to other deep-sea records, Shackleton, together with James Hays and Imbrie, were able to rigorously test the Milankovitch hypothesis — the idea that the great ice ages of recent Earth history were caused by subtle changes in the distribution of solar radiation across Earth's surface, in turn caused by orbital variations well known to astronomers.

They managed to use oxygen isotope and other records to detect the anticipated changes in orbital periodicities — of 19,000 and 23,000 years (precession), 41,000 years (obliquity) and 100,000 years (eccentricity) — in continuous sediment sequences. This provided the benchmark evidence that orbital changes act as a 'pacemaker' of climate change. Shackleton recognized that the orbital pacing of climate change made it possible to calibrate climate records from sedimentary sequences using the timing of the orbital cycles. Applying this principle, he gradually extended the orbitally tuned age-scale back some 30 million years. This has provided accurate dates for reversals of Earth's magnetic field, and for the evolution and extinction of marine organisms. It is a technique that has revolutionized the practice of stratigraphy.

Shackleton also pioneered the use and interpretation of carbon isotopes in palaeoclimate studies, an undertaking in which he moved on from studying the orbital forcing of glacial cycles to the positive feedbacks that amplify this forcing into dramatic changes in climate. He recognized that the carbon isotopic composition of the oceans is affected by the amount of carbon stored in forests and soils. And he was the first to use carbon isotope ratios in benthic foraminifera to assess the changing land reservoir of carbon between glacial and



UNIV. CAMBRIDGE

interglacial times. Shackleton also provided data that confirmed Wallace Broecker's proposal that the carbon isotopic difference between the surface and deep ocean could provide insights into the cause of the glacial–interglacial changes in carbon dioxide reconstructed from ice cores. Later, by combining deep-sea and ice-core records, he demonstrated that orbital eccentricity, the least well understood of the orbital forcings, probably affects climate through its influence on atmospheric carbon dioxide.

Shackleton, in his trademark sandals, was spirited and curiosity-driven. He let his students and an entire community share in his brilliance and vision. Through his meticulous work — he inspected every sample analysed in his lab — he was a champion of 'small science', showing by example that the most important data and the best ideas do not necessarily require high-priced enterprises.

It is revealing that Shackleton was an avid clarinetist, and taught a course on the physics of music at Cambridge. Music, more than any other art form, reveals itself through time and cyclic change. We owe a debt of gratitude to Nick Shackleton for his efforts to read the complex score of the Pleistocene climate symphony, and for helping to identify its composer — the variations in Earth's orbit and rotation. The challenge remains to reconstruct the orchestral players: the factors within the Earth system that work together to produce alternately the deep freeze of ice ages and the equable climates of interglacial periods, the last of which hosted the rise of human civilization.

Gerald H. Haug and Larry C. Peterson

Gerald H. Haug is at the Geoforschungszentrum Potsdam, 14473 Potsdam, Germany.
e-mail: haug@gfz-potsdam.de
Larry C. Peterson is at the Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida 33149, USA.
e-mail: lpeterson@rsmas.miami.edu

BRIEF COMMUNICATIONS

Spawning of eels near a seamount

Tiny transparent larvae of the Japanese eel collected in the open ocean reveal a strategic spawning site.

Discoveries of the larvae of the European and American eels, *Anguilla anguilla* and *A. rostrata*, in the Sargasso Sea^{1,2} and of the Japanese eel, *A. japonica*, in the Philippine Sea³ indicate that these freshwater eels migrate thousands of kilometres into the open ocean to spawn. Here we pinpoint a spawning location for Japanese eels after genetically identifying newly hatched larvae that we collected from the site. The restricted size of this spawning area ensures that the eel larvae enter a particular current that transports them to the freshwater areas in east Asia where they mature, and it also prevents them from being carried southwards away from their species range by a different local current.

It has been suggested that changes in oceanic conditions⁴ are contributing to the recent drastic decline of anguillid eels worldwide⁵, by disrupting their spawning areas and the transport of their larvae (leptocephali)⁴. But little is known about such spawning areas. On the basis of data collected on leptocephali over almost 50 years and analysis of their hatching dates, we have proposed that the Japanese eel

spawns near seamounts west of the Mariana Islands (14–17° N, 142–143° E), close to the time of the new moon⁶. We have now verified the location and timing of spawning by Japanese eels after collection and analysis of the newly hatched pre-leptocephali.

During research cruise KH-05-1 aboard *RV Hakuho Maru*, we collected and genetically identified 130 pre-leptocephali (size, 4.2–6.5 mm) and 60 leptocephali (11.7–18.4 mm) of the Japanese eel in the region of the North Equatorial Current in June 2005. The pre-leptocephali were collected around 14° N, 142° E, to the west of the Suruga Seamount in the southern part of the West Mariana Ridge (Fig. 1).

Some pre-leptocephali were identified as *A. japonica* on board by real-time polymerase chain reactions⁷; this identification was later confirmed by DNA sequencing of additional specimens⁸. (For details, see supplementary information.) The larvae were collected on the day of the new moon and for two days afterwards. They were at various stages of develop-

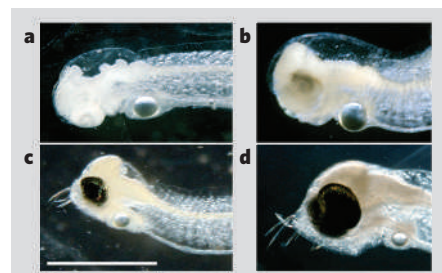


Figure 2 | Japanese eel pre-leptocephali at different stages of development. **a**, Head regions of early-stage larvae with no teeth, jaws or eye pigmentation; **b**, some larvae show early eye pigmentation; **c**, some have early teeth but no jaws; **d**, others have teeth that are more developed and early jaws. Larva lengths in **a–d** are 5.0, 4.7, 5.2 and 4.2 mm, respectively. Scale bar, 1 mm.

ment: some had unpigmented eyes and no teeth, others had pigmented eyes, early teeth and jaws (Fig. 2). Daily growth rings of calcium mineralization in 'ear stones' known as otoliths showed that the eels had hatched 2 to 5 days previously, indicating that spawning had occurred about 4 days before the new moon.

We have identified a precise spawning location by collecting newly hatched pre-leptocephali. The area seems to be much smaller than the spawning area estimated for the Atlantic eels (Fig. 1). It is located at an optimum latitude for the leptocephali to enter the Kuroshio Current that flows north towards the eels' habitat areas in east Asia. It seems that spawning occurs in a narrow range of latitudes because otherwise leptocephali would not find the northward flow and might instead enter the Mindanao Current, which flows southwards to places where there are no Japanese eels^{4,6}.

Katsumi Tsukamoto

Ocean Research Institute, University of Tokyo, Nakano, Tokyo 164-8639, Japan
e-mail: ktpc@ori.u-tokyo.ac.jp

- Schmidt, J. Smithson. *Inst. Annu. Rep.* **1924**, 279–316 (1925).
- McCleave, J. D., Kleckner, R. C. & Castonguay, M. *Am. Fish. Soc. Symp.* **1**, 286–297 (1987).
- Tsukamoto, K. *Nature* **356**, 789–791 (1992).
- Kimura, S., Inoue, T. & Sugimoto, T. *Fish. Oceanogr.* **10**, 51–60 (2001).
- Dekker, W. *Fisheries* **28**, 28–30 (2003).
- Tsukamoto, K. *et al. Environ. Biol. Fish.* **66**, 221–229 (2003).
- Watanabe, S. *et al. Mar. Biotechnol.* **6**, 566–574 (2004).
- Aoyama, J. *et al. Mar. Ecol. Progr. Ser.* **188**, 193–200 (1999).

Supplementary information accompanies this communication on Nature's website.

Received 18 November 2005; accepted 2 February 2006.

Competing financial interests: declared none.
doi:10.1038/439929a

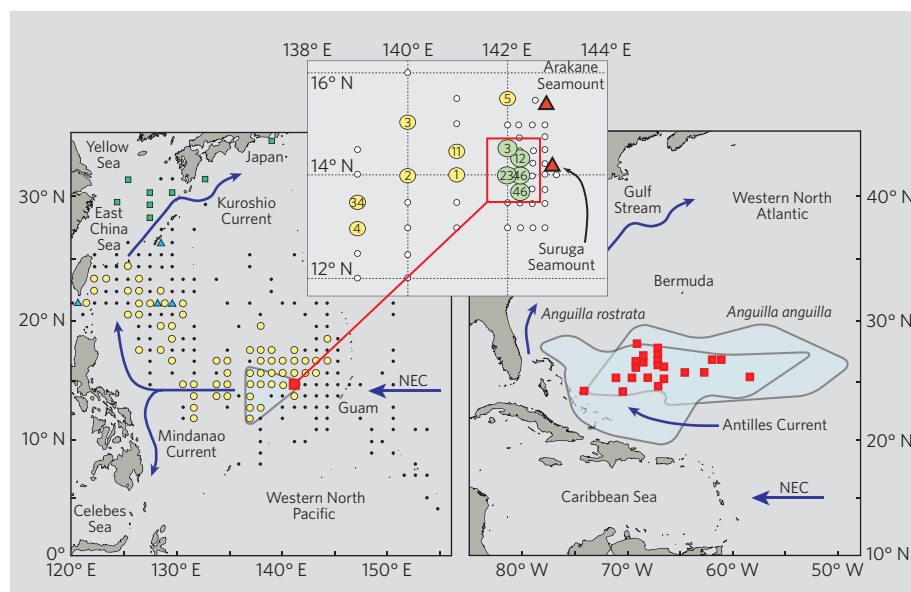


Figure 1 | Collection sites of small eel larvae. Left, catches of the Japanese eel, *Anguilla japonica*, in the western North Pacific Ocean between 1961 and 2002 ($n = 2,418$; ref. 6); yellow circles, sites where leptocephali larger than 7 mm were collected. Metamorphosing leptocephali (blue triangles) and oceanic glass eels (green squares) were also found. Black dots, sites where no larvae were found. Right, eel catches in the western North Atlantic Ocean (1913–85; ref. 2). Shaded areas, sites where leptocephali (≤ 10 mm in size) of each species were collected². Red squares mark collection sites of pre-leptocephali (≤ 7 mm) of the Japanese eel (left) or of both Atlantic eel species² (right). NEC, North Equatorial Current. Inset, distribution and number of Japanese eel pre-leptocephali (≤ 7 mm; green circles) and leptocephali (> 7 mm; yellow circles) collected during the time of the new moon in June 2005. Red triangles, seamounts. White circles, stations where no Japanese eel leptocephali were found.

REVIEWS

A new radiocarbon revolution and the dispersal of modern humans in Eurasia

Paul Mellars¹

Radiocarbon dating has been fundamental to the study of human cultural and biological development over the past 50,000 yr. Two recent developments in the methodology of radiocarbon dating show that the speed of colonization of Europe by modern human populations was more rapid than previously believed, and that their period of coexistence with the preceding Neanderthal was shorter.

Radiocarbon dating, first developed by W. F. Libby in 1947, works on the assumption that the proportion of the radioactively unstable ^{14}C isotope to stable ^{12}C has remained effectively constant and homogeneous in the Earth's atmosphere over the past 50,000 yr (the effective limit of the method), and that the rate of decay of the ^{14}C isotope (with a half-life of approximately 5,730 yr) can be used as a measure of the age of all forms of once-living materials (principally plant and animal remains) that secured their original carbon content directly or indirectly from the contemporaneous atmosphere¹. The degree of precision in dating declines with the increasing age of the samples, but even so, with the use of high-precision accelerator mass spectrometer techniques this method can produce dates with an accuracy of a few hundred years extending back to at least 45,000 yr before present (BP)². Because this is better than the precision that can be obtained by almost any other dating methods in this time range (with the exception of some uranium decay methods, only applicable to carbonate formations¹), radiocarbon continues to provide the central timescale for all of the current studies of the geographical dispersal and cultural development of early anatomically and behaviourally modern human populations (that is, *Homo sapiens*) after their initial dispersal from Africa between approximately 50,000 and 60,000 yr ago^{3–7}.

The application of radiocarbon dating to these crucial early phases in modern human development has, however, been critically dependent on two potential sources of error in the accuracy of radiocarbon age estimates. The first is the impact of even minuscule quantities of contamination by more recent, intrusive carbon into the dated samples (Fig. 1a). This can be illustrated by the fact that contamination by only 1% of modern carbon in a sample actually 40,000 yr in age would reduce the measured age of the sample by over 7,000 yr—an effect that doubles with every additional half-life (5,730 yr) in the age of the sample^{1,8,9}. The second is the long-established recognition that the original proportion of ^{14}C to ^{12}C in the Earth's atmosphere has not remained constant over the past 50,000 yr, but has diverged sharply from present-day values, principally due to past variations in the intensity of the Earth's magnetic field and the shorter-term effects of sunspots on the amount of cosmic radiation reaching the upper atmosphere^{1,10–17}. The combination of these two sources of potential error in radiocarbon dating has been a major complication for archaeologists and palaeoanthropologists attempting to unravel the true patterns of evolutionary and behavioural development during these early phases of modern human development^{5,8,9}.

Here I show that recent developments in the methodology of radiocarbon dating have had a dramatic impact on both of these sources of error. New developments in the preparation of bone samples would seem to have largely resolved the problem of modern contamination in the dating of these samples. And a spate of new data on the calibration of radiocarbon dates into 'absolute' calendar years show that the speed of dispersal of anatomically modern populations across Europe was much more rapid than previously believed.

New developments in radiocarbon dating

Two recent developments, in particular, have effectively revolutionized the application of radiocarbon dating to the study of modern human origins and dispersal in Eurasia. In the first place, recent developments in the pre-treatment of bone samples at the University of Oxford⁸ have led to radical improvements in the procedures for the effective purification of bone collagen to eliminate contamination by more recent carbon—especially in the case of older bone samples, which have always provided the most widely available materials for dating from most early modern human sites^{8,9,18}. The new techniques involve the 'ultrafiltration' of the prepared gelatin samples to separate out the smaller and lower molecular weight fractions, which seem to have been the major source of more recent organic contaminants (from percolating humic acids, organic salts, heavily degraded collagen, and so on) in the dated samples^{8,18}. Recent applications of this procedure to a range of samples that had previously been processed by means of conventional pre-treatment techniques have led to dates that are frequently between 2,000 and 7,000 yr older than the original age estimates¹⁸ (see Fig. 1b). For example, dating of a later Aurignacian bone point from Uphill quarry, Somerset, which had previously been dated to $28,080 \pm 360$ yr BP produced a revised date of $31,730 \pm 250$ yr BP, whereas the re-dating of a rhinoceros bone from the site of Kent's cavern, Devonshire, increased the measured age from $30,220 \pm 460$ to $37,200 \pm 550$ yr BP.

The second breakthrough has emerged from recent research into the fluctuation patterns of the original ^{14}C content of the Earth's atmosphere over the past 50,000 yr^{10–17}. The most significant and internally consistent results have come from the dating of a series of 280 stratified radiocarbon samples recovered from a long sequence of deep-sea sediments in the Cariaco Basin near Venezuela, dated in 'calendrical' terms by reference to closely matching patterns of oxygen isotope ($^{18}\text{O}/^{16}\text{O}$) fluctuations in the independently dated Greenland Ice Sheet Project 2 (GISP2) ice-core records from central

¹Department of Archaeology, Cambridge University, Cambridge CB2 3DZ, UK.

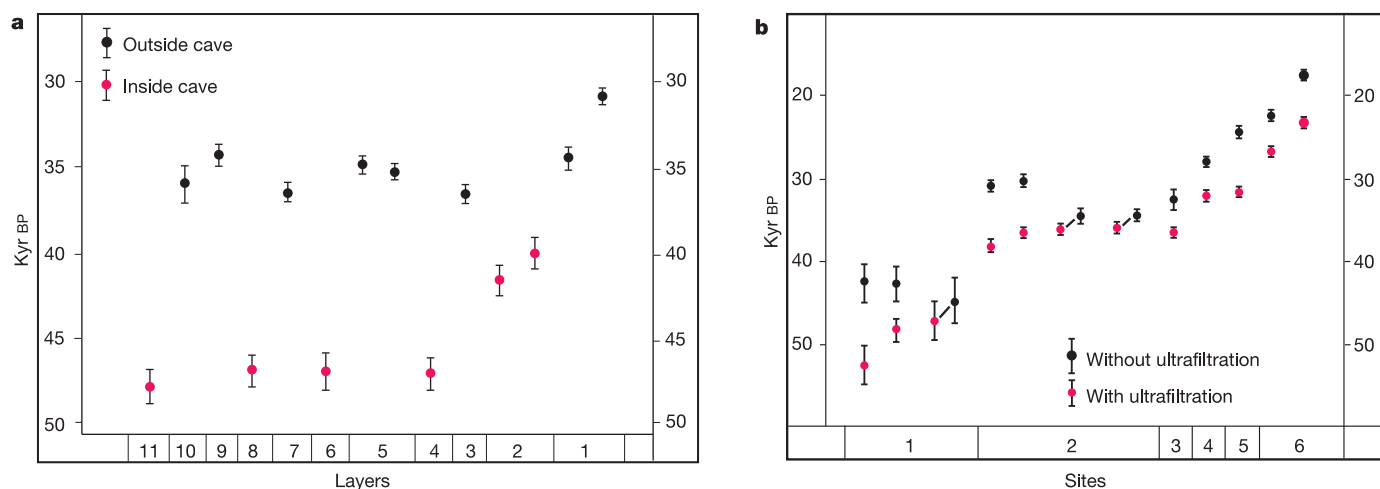


Figure 1 | Contamination effects in radiocarbon dating. **a**, Contamination effects on radiocarbon samples from the Sesselfelsgrötte cave (south Germany). The graph compares dates for successive levels based on bone samples collected from the exterior versus interior areas of the cave, dated by the Groningen radiocarbon accelerator laboratory. The samples from the unprotected exterior part of the cave have clearly been heavily affected by continuous percolation of groundwater containing humic acids and other organic contaminants from the present surface. The effects of the contamination are equivalent to approximately 1% of contamination by

modern carbon in all of these samples. Data from ref. 48. **b**, The effects of ultrafiltration preparation techniques on the dating of bone samples. The graph compares dates on the same bone samples made respectively with and without the use of ultrafiltration pre-treatment techniques at the Oxford radiocarbon laboratory^{8,18}. The sites are: 1, Pinhole cave (Derbyshire); 2, Kent's cavern (Devonshire); 3, Brixham cave (Devon); 4, Uphill cave (Somerset); 5, Hyæna den (Somerset); 6, Paviland cave (Swansea). Data reproduced with permission from Jacobi, Higham and Bronk Ramsey (ref. 18).

Greenland¹⁶; from a similar sequence of datings from deep-sea sediments adjacent to the Iberian coast¹⁷; from a series of 152 paired radiocarbon and high-precision uranium/thorium (U/Th) measurements on a number of fossil coral formations from the tropical Atlantic and Pacific¹⁵; and from a sequence of similar, combined ¹⁴C and U/Th dating of a long cave stalagmite formation from the island of Socotra off the Arabian coast¹⁷. The results of these different measurements have recently been compared to give a “best-estimation” comparison between measured radiocarbon ages and ‘absolute’ calendar ages over the past 50,000 yr, in the recently published NotCal04 calibration study presented at the 2003 radiocarbon calibration conference in Wellington, New Zealand¹⁷ (Fig. 2).

Although the results of the separate calibration records differ in certain respects, two significant patterns have emerged from these correlations. First, as noted above, we can now see that radiocarbon dates diverge sharply from true ages in the time range from about 10,000 to 45,000 yr BP, apparently due principally to the effects of the major Laschamp and Lake Mono geomagnetic excursion events at around 41,000 and 28,000 calendar years ago, respectively, which sharply increased the amount of cosmic radiation reaching the Earth's upper atmosphere and accordingly increased the ¹⁴C content of the atmosphere^{13,14,16,19}. The new calibration curves (Fig. 2) show that a measured radiocarbon date of 40,000 yr BP translates into an actual (calendar) date of approximately 43,000 yr BP, whereas a radiocarbon date of 35,000 yr BP translates into a calendar age of about 40,500 yr BP. The systematic displacement of radiocarbon ages from true calendar ages has obvious implications for any comparison between dates for archaeological or geological sites produced by means of radiocarbon as compared to other dating methods, such as uranium/thorium or thermoluminescence¹. The new calibration curves also reveal that the observed pattern of deviations between radiocarbon ages and real ages within this time range follows a relatively simple, smooth pattern, apparently without any of the sudden and aberrant oscillations in the atmospheric ¹⁴C content that had been claimed from some earlier attempts at calibration based on studies of stratified lake sediments at Lake Suigetsu in Japan¹¹ and studies of a cave stalagmite formation in the Bahamas¹² (Fig. 2). This is clearly a highly important discovery, as the occurrence of erratic

oscillations of this kind would inevitably lead to equally erratic fluctuations in measured radiocarbon ages, and in some cases to possible major ‘plateaux’ or even substantial reversals in the recorded patterns of radiocarbon dates.

There is another important methodological implication of this new radiocarbon calibration. As seen in Fig. 2, the effect of the new calibration is to reveal a rapid change in measured radiocarbon ages over the period between about 40,000 and 35,000 radiocarbon years ago, due to the rapid changes in the atmospheric ¹⁴C content over this time range. In practice, this means that the radiocarbon dates that cover an apparent (that is, measured) period of 5,000 yr in reality cover an actual time range of only around 3,000 yr. In practical terms this means that the relative precision and chronological resolution of radiocarbon measurements over this time range is significantly greater than that in other, adjacent parts of the radiocarbon time-scale, where the atmospheric radiocarbon content followed a more regular, linear pattern. Thus, what had previously seemed to be a

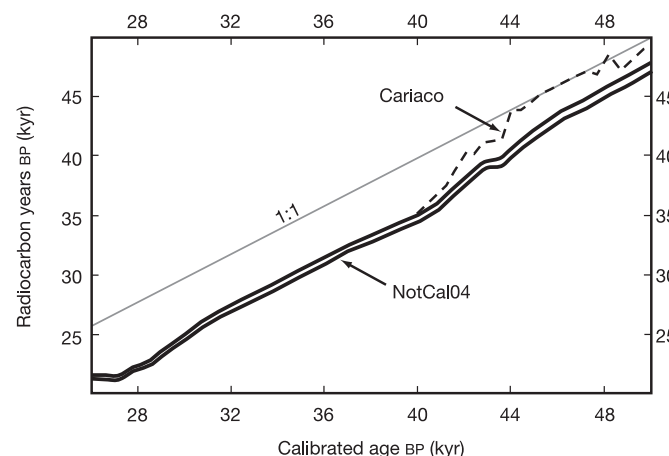


Figure 2 | Radiocarbon calibration curves for the 25,000–50,000-yr time range. The graph shows the recent NotCal04 “best estimation” calibration curve of ref. 17 (adjacent black lines) and the alternative calibration in the 40,000–50,000-yr range based on the recent Cariaco Basin data¹⁶.

'nightmare' period for the application of radiocarbon dating in early prehistory now emerges as the period when the method is apparently performing at its best. Combined with the improvements in the methods for the pre-treatment and purification of bone samples, this is good news indeed.

Palaeoanthropological implications

The central question is exactly what implications do these new developments in radiocarbon dating have for our understanding of the patterns of human development during the critical 50,000–35,000 yr BP time range when we know that anatomically and behaviourally modern populations were expanding across western Eurasia from their original African homeland^{3–7,20,21}.

Overall, perhaps the most significant impact of this new radiocarbon calibration lies in its implications for the relative speed with which anatomically and behaviourally modern human populations expanded across Eurasia, and the extent to which they overlapped with the preceding Neanderthal populations within the different regions. On the basis of discoveries of fossilized human skeletal remains and a number of long stratified archaeological sequences in Israel and Lebanon it has been recognized that populations of anatomically and behaviourally modern humans, equipped with

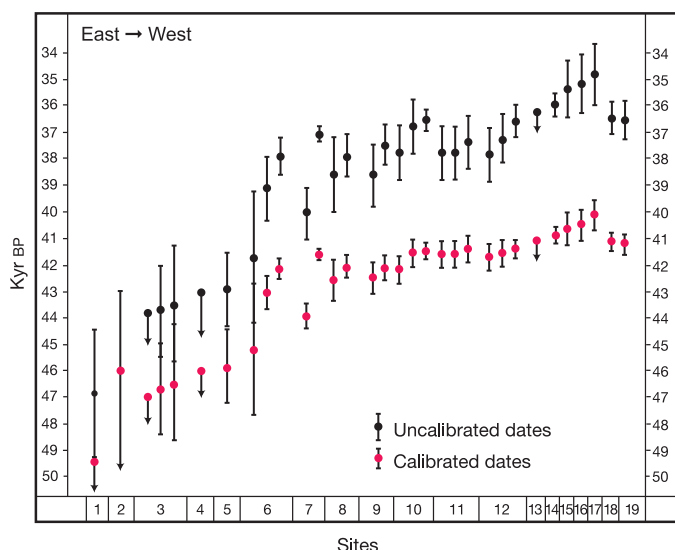


Figure 3 | Comparison of calibrated and uncalibrated radiocarbon dates for the dispersal of modern humans across Europe and the Near East. The numbers, arranged from east to west, refer to archaeological sites belonging to the Emiran, Ahmarian, Bacho-Kirian, Bohunician, Proto-Aurignacian and Aurignacian technologies in different regions, all believed to be the products of early anatomically modern populations^{4,5,22,23}. Calibrated dates are based on the mid-points of the NotCal04 calibration shown in Fig. 2. Application of the Cariaco Basin calibration curve (Fig. 2) would yield substantially younger calibrated ages for the 40,000–50,000-yr range, and a correspondingly faster rate of dispersal across Europe. Owing to the slope of the calibration curves, the error bars (± 1 s.d.) on the calibrated dates are smaller than those on the uncalibrated dates. Only the oldest radiocarbon measurements are plotted from each region, on the assumption that these are likely to be least affected by contamination with more recent carbon. The sites plotted are: 1, Boker Tachtit (Israel); 2, Ksar Akil (Lebanon); 3, Kebara (Israel); 4, Bacho Kiro (Bulgaria); 5, Bohunice (Czech Republic); 6, Willendorf (Austria); 7, Grotta Fumane (Italy); 8, El Pina (Italy); 9, Keilberg-Kirche (Germany); 10, Geissenklösterle (Germany); 11, L'Arbreda (Spain); 12, Abric Romani (Spain); 13, Châtelperron (France); 14, La Rochette (France); 15, Abri Caminade (France); 16, Abri Castanet (France); 17, Roc de Combe (France); 18, Isturitz (France); 19, Cueva Morín (Spain) (refs 4, 28, 30, 33, 47, 49–56). The date range shown for Ksar Akil (site 2) is based on age/depth extrapolations from overlying radiocarbon measurements⁵⁵. The currently controversial dates for El Castillo (Spain), with disputed archaeological and skeletal associations⁵⁶, have been omitted.

typically Upper Palaeolithic technology, had appeared in the near eastern region by at least 45,000 yr BP (in uncalibrated radiocarbon terms), and had apparently spread to parts of southeastern Europe shortly after this time^{4,5,20,22,23}. However, the dispersal of these populations across the remaining areas of central and western Europe seemed, on the basis of the 'raw' radiocarbon dates, to have taken a period of at least a further 7,000 yr, between approximately 43,000 and 36,000 yr in uncalibrated radiocarbon terms. This implies an overall rate of dispersal of these populations of around 0.3 km yr^{-1} . In the light of the new calibration data (Fig. 2) we can now see that this period compresses to an actual time span of only about 5,000 yr (from about 46,000 to 41,000 yr BP in calibrated terms), implying a substantially faster rate of dispersal across this region (see Fig. 3). If we were to adopt the newly published Cariaco Basin calibration curve (Fig. 2), the result would be an even faster migration rate of around 0.4 km yr^{-1} . This rate of dispersal is broadly similar to the later dispersal of early agricultural communities across the same geographical range, between about 10,000 and 6,000 (calibrated) years BP²⁴. It is equally interesting to see that the two dispersals seem to have followed closely similar geographical routes: one along the Mediterranean coast from Israel to northern Spain and the other along the Danube valley from the Balkans to southern Germany and eventually western France^{5,21–23} (Fig. 4). The rapid spread of the early modern human populations was probably facilitated by a major improvement in climatic conditions in Europe between about 43,000–41,000 yr (calibrated) BP (the period of the Hengelo interstadial), which would inevitably have made a process of population expansion from southeast to northwest across Europe easier to achieve^{14,23,25,26}. Climatic modelling studies suggest that both summer and winter temperature isotherms shifted by around 1,000 km westwards during this interval²⁷, closely paralleling the westwards expansion of the earliest anatomically modern populations from central Europe to western France.

The same chronological pattern points to a substantially shorter period of chronological and demographic overlap between the earliest intrusive populations of anatomically and behaviourally modern humans and the last survivors of the preceding Neanderthal populations within the different regions of Europe. Although this has often been estimated in the region of approximately 10,000 yr within Europe as a whole^{3,28,29}, we can now see from the new calibrated

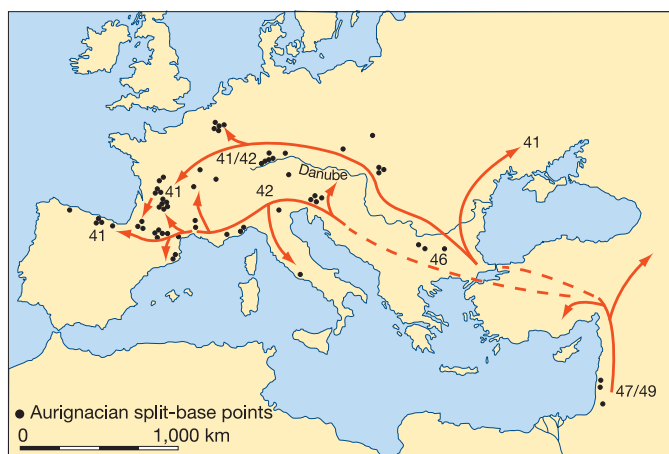


Figure 4 | Dispersal routes of modern human populations across Europe. The dates shown for each region (in thousands of yr BP; range 47,000–41,000) are based on the calibrated (that is, calendrical age) radiocarbon measurements plotted in Fig. 3, derived from the NotCal04 calibration data shown in Fig. 2. The distribution of classical Aurignacian split-base bone/antler points is also shown for comparison, although these are not necessarily associated directly with the adjacent age estimates. Note the contrast between these dates and the uncalibrated radiocarbon ages plotted in ref. 5.

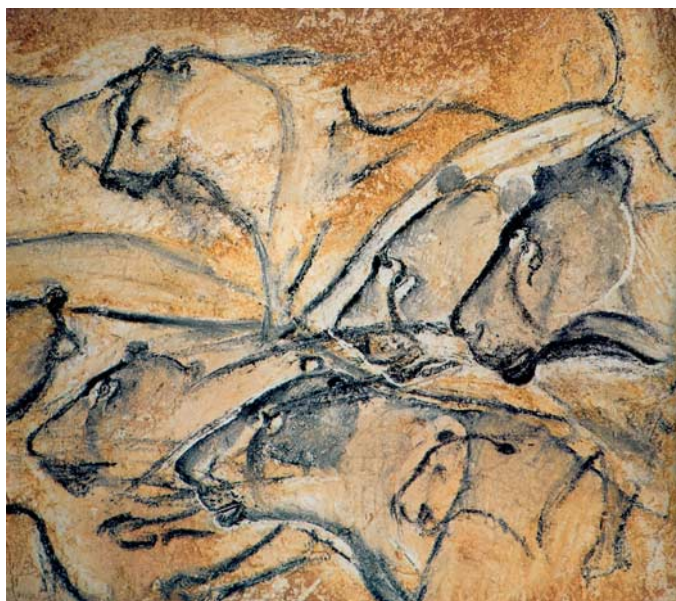


Figure 5 | Drawing of heads of cave lions in the Chauvet cave (south-east France). The drawings in the Chauvet cave were originally dated by radiocarbon dating of the charcoal used to produce the drawings to around 31,000–32,000 yr BP, but are now redated by the new NotCal04 calibration (Fig. 2) to around 36,000 yr BP in calendrical terms. Reproduced from Clottes (ref. 32) with permission (see Acknowledgements).

chronology that this must be shortened to at most about 6,000 yr (at least in the more central and northern parts of Europe), with periods of overlap within the individual regions of Europe (such as western France) of perhaps only 1,000–2,000 yr^{28,30}. Evidently the native Neanderthal populations of Europe succumbed much more rapidly to competition from the expanding biologically and behaviourally modern populations than previous estimates have generally assumed.

We can now see from the new calibration data that some of the most striking cultural achievements of the new anatomically modern human populations in Europe appeared much earlier than the original radiocarbon dates had suggested. As Bard *et al.*³¹ have pointed out, the spectacular cave art in the Chauvet cave³² in southeastern France (Fig. 5) can now be seen to date from around 36,000 yr BP in real terms (as compared to about 31,000–32,000 yr BP in radiocarbon terms), whereas the equally dramatic eruption of elaborate bone, antler and ivory technology, and the associated proliferation of various forms of personal ornaments, appeared in association with the earliest Aurignacian populations in central and western Europe by at least 41,000 yr BP in calendar terms^{5,33–35}. Whether or not these innovations coincided with the first appearance of fully structured language and associated complex social organization among the earliest biologically modern human populations in Europe is still disputed^{35–38}, but there can be little doubt that these new cultural and behavioural developments had a critical role in the rapid replacement of the European Neanderthals by the new, intrusive, biologically modern populations from an ultimately African source.

There are increasing indications that over many areas of Europe the final demise of the Neanderthal populations may have coincided with the sudden onset of very much colder and drier climatic conditions marked in the deep-sea records by the period of Heinrich event 4 (dated to about 35,000 yr BP in radiocarbon terms, or about 40,500 yr BP in calibrated terms), during which many icebergs broke off from the North Atlantic ice sheet and sharply depressed sea and land temperatures over large areas of Europe^{26,39–41}. If, as most of the current evidence suggests, the new anatomically modern human populations were better equipped technologically and culturally to

deal with these severe glacial conditions^{30,35,41,42}, then this could have delivered the *coup de grace* to the Neanderthals in many parts of western and central Europe, in their economic and demographic competition with the incoming modern groups^{26,30}. Whether some of the Neanderthal groups may have survived longer in some of the more southerly areas of Europe, such as the Iberian peninsula or the Balkans, is more controversial^{3,29,40,43–45}, and once again depends heavily on the reliability of the available radiocarbon dates used to support these apparently late Neanderthal survival events.

Future prospects

Although these are all major breakthroughs, we should continue to be cautious about the current state of radiocarbon dating in human prehistory. Palaeolithic archaeology has inherited a massive legacy of published radiocarbon dates accumulated over the past 40 yr, a high proportion of which are almost certainly serious underestimates of the true ages of the samples, as a direct result of the major contamination effects discussed earlier^{8–18} (Fig. 1). This is almost certainly the explanation for the long tail of dates for late Mousterian (that is Neanderthal) sites in Europe extending long after 35,000 yr BP^{9,29}, and the similar tail of Aurignacian dates extending after 30,000 yr BP—far into the range of the stratigraphically younger Gravettian sites^{22,29,33}. This is also the explanation for the successive datings of the early anatomically modern human skeleton from the Paviland cave (Swansea) becoming progressively younger from (initially) around 18,000 yr BP to (currently) around 26,000 yr BP⁴⁶ and for the large differences in the measured ages of radiocarbon samples collected respectively from the interior versus exterior areas of cave sites such as the Grotta Fumane (Italy)⁴⁷, Sesselfelsgrötte (Germany)⁴⁸ and elsewhere. These show the dramatic effects of contamination by percolating groundwater in samples collected from the ‘wet’ versus ‘dry’ areas of these sites. The results of the long series of ¹⁴C measurements from Sesselfelsgrötte provide, in effect, a 40,000-yr-long laboratory experiment in the effects of percolating groundwater on the contamination of radiocarbon samples, leading, in this case, to errors of between 5,000 and 12,000 yr in the measured ages of the samples collected from the exposed, exterior parts of the cave (see Fig. 1a). For the same reason we must be extremely cautious about accepting the published dates for critical discoveries such as the late Neanderthal fossils from the Zafarraya cave (Spain) or from Vindija (Croatia)—both dated to between about 31,000 and 28,000 yr BP—at face value^{43–45}.

We should bear in mind that all of the recent attempts at radiocarbon calibration within the 25,000–50,000-yr time range are still provisional, and potentially subject to a number of errors in both the radiocarbon and associated calendrical age estimates of the dated samples^{13–17}. A final, definitive calibration curve for this time range will depend on the results of new calibration studies, at present being pursued in several different laboratories¹⁷. The full implications of these studies for the interpretation of the human archaeological and evolutionary record will need to be kept under active and vigilant review.

1. Aitken, M. J. *Science-based Dating in Archaeology* (Longmans, London, 1990).
2. Bronk Ramsey, C., Higham, T. & Leach, P. Towards high-precision AMS: progress and limitations. *Radiocarbon* **46**, 17–24 (2004).
3. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357**, 563–579 (2002).
4. Bar-Yosef, O. In *The Geography of Neandertals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 107–156 (Peabody Museum of Harvard University, Peabody Museum Bulletin 8, Cambridge, Massachusetts, 2000).
5. Mellars, P. Neanderthals and the modern human colonization of Europe. *Nature* **432**, 461–465 (2004).
6. Forster, P. Ice ages and the mitochondrial DNA chronology of human dispersals: a review. *Phil. Trans. R. Soc. Lond. B* **359**, 255–264 (2004).
7. Forster, P. & Matsumura, S. Did early humans go north or south? *Science* **308**, 965–966 (2005).
8. Bronk Ramsey, C., Higham, T., Bowles, A. & Hedges, R. Improvements to the pretreatment of bone at Oxford. *Radiocarbon* **46**, 155–163 (2004).

9. Jöris, O., Alvarez Fernández, E. & Weninger, B. Radiocarbon evidence of the Middle to Upper Palaeolithic transition in southwestern Europe. *Trabajos Prehistoria* **60**, 15–38 (2003).
10. Voelker, A. H. L. *et al.* Correlation of marine ^{14}C ages from the Nordic seas with the GISP2 isotope record: implications for ^{14}C calibration beyond 25 ka BP. *Radiocarbon* **40**, 517–534 (1998).
11. Kitagawa, H. & van der Plicht, J. Atmospheric radiocarbon calibration to 45,000 yr BP. Late Glacial fluctuations and cosmogenic isotope production. *Science* **279**, 1187–1190 (1998).
12. Beck, J. W. *et al.* Extremely large variations of atmospheric ^{14}C concentration during the last glacial period. *Science* **292**, 2453–2458 (2001).
13. Bard, E., Ménot-Combes, G. & Rostek, F. Present status of radiocarbon calibration and comparison records based on Polynesian corals and Iberian Margin sediments. *Radiocarbon* **46**, 1189–1202 (2004).
14. Shackleton, N. J., Fairbanks, R. G., Chiu, T. & Parrenin, F. Absolute calibration of the Greenland time scale: implications for Antarctic time scales and for delta ^{14}C . *Quat. Sci. Rev.* **23**, 1513–1522 (2004).
15. Fairbanks, R. G. *et al.* Radiocarbon calibration curve spanning 0 to 50,000 years BP based on paired $^{230}\text{Th}/^{234}\text{U}/^{238}\text{U}$ and ^{14}C dates on pristine corals. *Quat. Sci. Rev.* **24**, 1781–1796 (2005).
16. Hughen, K. *et al.* ^{14}C activity and global carbon cycle changes over the past 50,000 years. *Science* **303**, 202–207 (2004).
17. van der Plicht, J. *et al.* NotCal04-comparison/calibration ^{14}C records 26–50 cal. kyr BP. *Radiocarbon* **46**, 1225–1238 (2004).
18. Jacobi, R. M., Higham, T. F. G. & Bronk Ramsey, C. Improving the AMS radiocarbon dating of Middle and Upper Palaeolithic bone in the British Isles. *J. Quat. Sci.* (in the press).
19. Laj, C. *et al.* Geomagnetic field intensity, North Atlantic deep water and atmospheric delta ^{14}C during the last 50 kyr. *Earth Planet. Sci. Lett.* **200**, 177–190 (2002).
20. Mellars, P. A. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. *Phil. Trans. R. Soc. Lond. B* **337**, 225–234 (1992).
21. Mellars, P. A. in *The Peopling of Britain: the Shaping of a Human Landscape* (eds Slack, P. & Ward, R.) 39–67 (Oxford Univ. Press, Oxford, 2001).
22. Kozłowski, J. K. & Otte, M. The formation of the Aurignacian in Europe. *J. Anthropol. Res.* **56**, 513–534 (2000).
23. Mellars, P. A. Archaeology and the dispersal of modern humans in Europe. *Evol. Anthropol.* (in the press).
24. Whittle, A. in *Prehistoric Europe: an Illustrated History* (ed. Cunliffe, B. W.) 136–166 (Oxford Univ. Press, Oxford, 1998).
25. Mellars, P. in *The Lower and Middle Palaeolithic: Colloquia* (eds Bar-Yosef, O., Cavalli-Sforza, L. L., March, R. J. & Piperno, M.) 225–237 (ABACO editions, Forlì, 1996).
26. Mellars, P. A. in *Neanderthals and Modern Humans in Western Asia* (eds Akazawa, T., Aoki, K. & Bar-Yosef, O.) 493–508 (Plenum, New York, 1998).
27. Barron, E., van Andel, T. & Pollard, D. in *Neanderthals and Modern Humans in the European Landscape During the Last Glaciation* (eds van Andel, T. H. & Davies, W.) 57–68 (McDonald Institute for Archaeological Research, Cambridge, 2003).
28. Mellars, P. A. in *The Geography of Neanderthals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 493–507 (Peabody Museum of Harvard University, Peabody Museum Bulletin 8, Cambridge, Massachusetts, 2000).
29. van Andel, T. H., Davies, W. & Weninger, B. in *Neanderthals and Modern Humans in the European Landscape during the Last Glaciation* (eds van Andel, T. H. & Davies, W.) 31–56 (McDonald Institute for Archaeological Research, Cambridge, 2003).
30. Gravina, B., Mellars, P. & Bronk Ramsey, C. Radiocarbon dating of interstratified Neanderthal and early modern human occupations at the Chatelperronian type-site. *Nature* **438**, 51–56 (2005).
31. Bard, E., Rostek, F. & Ménot-Combes, G. A better radiocarbon clock. *Science* **303**, 178–179 (2004).
32. Clottes, J. (ed.) *La Grotte Chauvet: l'art des Origines* (Seuil, Paris, 2001).
33. Conard, N. J. & Bolus, M. Radiocarbon dating the appearance of modern humans and timing of cultural innovations in Europe: new results and new challenges. *J. Hum. Evol.* **44**, 331–371 (2003).
34. White, R. in *Before Lascaux: the Complex Record of the Early Upper Paleolithic* (eds Knecht, H., Pike-Tay, A. & White, R.) 277–300 (CRC Press, Boca Raton, 1993).
35. Gamble, C. *The Palaeolithic Societies of Europe* (Cambridge Univ. Press, Cambridge, 1999).
36. Mithen, S. *The Prehistory of the Mind* (Thames & Hudson, London, 1996).
37. Lewis-Williams, D. *The Mind in the Cave* (Thames & Hudson, London, 2002).
38. Bickerton, D. *Language and Human Behavior* (Univ. Washington Press, Seattle, 1995).
39. Shackleton, N. J., Hall, M. A. & Vincent, E. Phase relationships between millennial-scale events 64,000–24,000 years ago. *Paleoceanography* **15**, 565–569 (2000).
40. d'Errico, F. & Sánchez Goñi, M. F. Neandertal extinction and the millennial scale climatic variability of OIS 3. *Quat. Sci. Rev.* **22**, 769–788 (2003).
41. Stringer, C. *et al.* in *Neanderthals and Modern Humans in the European Landscape during the Last Glaciation* (eds van Andel, T. H. & Davies, W.) 233–240 (McDonald Institute for Archaeological Research, Cambridge, 2003).
42. Aiello, L. C. & Wheeler, P. in *Neanderthals and Modern Humans in the European Landscape during the Last Glaciation* (eds van Andel, T. H. & Davies, W.) 147–166 (McDonald Institute for Archaeological Research, Cambridge, 2003).
43. Hublin, J.-J. in *The Geography of Neanderthals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 157–182 (Peabody Museum of Harvard University, Peabody Museum Bulletin 8, Cambridge, Massachusetts, 2000).
44. Carbonell, E., Vaquero, M., Maroto, J., Rando, J. M. & Mallol, C. in *The Geography of Neanderthals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 5–34 (Peabody Museum of Harvard University, Peabody Museum Bulletin 8, Cambridge, Massachusetts, 2000).
45. Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanić, I. & Paunović, M. Direct radiocarbon dates for Vindija G1 and Velika Pećina Late Pleistocene hominid remains. *Proc. Natl Acad. Sci. USA* **96**, 12281–12286 (1999).
46. Aldhouse-Green, S. (ed.) *Paviland Cave and the "Red Lady": a Definitive Report* (Western Academic and Specialist Press, Bristol, 2000).
47. Bartolomei, G. *et al.* La Grotte de Fumane. Un site aurignacien au pied des Alpes. *Preistoria Alpina* **28**, 131–179 (1992).
48. Richter, J. Die ^{14}C -daten aus der Sesselfelsgrötte und die zeitstellung Micoquien/M.M.O. *Germania* **80**, 1–22 (2002).
49. Bergman, C. A. & Stringer, C. B. Fifty years after: Egbert, an early Upper Palaeolithic juvenile from Ksar Akil, Lebanon. *Paléorient* **15**, 99–111 (1989).
50. Maíllo Fernández, J. F., Valladas, H., Cabrera Valdés, V. & Bernaldo de Quiros, F. Nuevos dataciones para el Paleolítico superior de Cueva Morin (Villanueva de Villascusa, Cantabria). *Espacio, Tiempo y Forma Ser. 1. Prehistoria Arqueología* **14**, 145–150 (2001).
51. Haesaerts, P. & Teyssandier, N. in *The Chronology of the Aurignacian and of the Transitional Complexes: Dating, Stratigraphies, Cultural Implications* (eds Zilhao, J. & d'Errico, F.) 133–151 (Instituto Português de Arqueologia, Trabalhos de Arqueologia 33, Lisbon, 2003).
52. Kuhn, S. L. & Bietti, A. in *The Geography of Neanderthals and Modern Humans in Europe and the Greater Mediterranean* (eds Bar-Yosef, O. & Pilbeam, D.) 49–76 (Peabody Museum of Harvard University, Peabody Museum Bulletin 8, Cambridge, Massachusetts, 2000).
53. Le Bon, F. *L'Aurignacien Entre Mer et Océan. Réflexions sur l'unité des Phases Anciennes de l'Aurignacien dans le Sud de la France* Mémoire 29 (Société Préhistorique de France, Paris, 2002).
54. Uthmeier, T. Ein bemerkenswert frühes inventar des Aurignacien von freilandfundstelle "Keilberg-Kirche" bei Regensburg. *Archäologisches Korrespondenzblatt* **26**, 233–248 (1996).
55. Mellars, P. & Tixier, J. Radiocarbon-accelerator dating of Ksar 'Aqil (Lebanon) and the chronology of the Upper Palaeolithic sequence in the Middle East. *Antiquity* **63**, 761–768 (1989).
56. Zilhao, J. & d'Errico, F. in *The Chronology of the Aurignacian and of the Transitional Complexes: Dating, Stratigraphies, Cultural Implications* (eds Zilhao, J. & d'Errico, F.) 313–349 (Instituto Português de Arqueologia, Trabalhos de Arqueologia 33, Lisbon, 2003).

Acknowledgements I thank N. Shackleton, C. Bronk Ramsey, T. Higham, R. Hedges, H. Valladas, N. Mercier, M. Fontugne, N. Tisnérat-Laborde, M. Paterne, B. Weninger, W. Davies, O. Bar-Yosef, J. Richter and J. Kozłowski for discussions of the points raised in the paper. Figure 5 was reproduced with permission from the French Ministry of Culture and Communication, Regional Direction for Cultural Affairs-Rhône-Alpes, Regional Department of Archeology (image 20).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to P.M. (pam59@cam.ac.uk).

ARTICLES

Adaptive filtering enhances information transmission in visual cortex

Tatyana O. Sharpee^{1,2}, Hiroki Sugihara², Andrei V. Kurgansky², Sergei P. Rebrik², Michael P. Stryker^{1,2} & Kenneth D. Miller^{1,2,3}

Sensory neuroscience seeks to understand how the brain encodes natural environments. However, neural coding has largely been studied using simplified stimuli. In order to assess whether the brain's coding strategy depends on the stimulus ensemble, we apply a new information-theoretic method that allows unbiased calculation of neural filters (receptive fields) from responses to natural scenes or other complex signals with strong multipoint correlations. In the cat primary visual cortex we compare responses to natural inputs with those to noise inputs matched for luminance and contrast. We find that neural filters adaptively change with the input ensemble so as to increase the information carried by the neural response about the filtered stimulus. Adaptation affects the spatial frequency composition of the filter, enhancing sensitivity to under-represented frequencies in agreement with optimal encoding arguments. Adaptation occurs over 40 s to many minutes, longer than most previously reported forms of adaptation.

The neural circuits in the brain that underlie our behaviour are well suited for processing of real-world—or natural—stimuli. These neural circuits, especially at the higher stages of neural processing, may be largely or completely unresponsive to many artificial stimulus sets used to analyse the early stages of sensory processing and, more generally, for systems analysis. Thus, natural stimuli may be necessary to study higher-level neurons. Characterizing neural responses to natural stimuli at early or intermediate stages of neural processing, such as the primary visual cortex, is a necessary step for systematic studies of higher-level neurons. Neural responses are also known to be highly nonlinear^{1–3} and adaptive^{4–20}, making them difficult to predict across different stimulus sets²¹. Therefore, even early in visual processing, characterizations based on simplified stimuli may not be adequate to understand responses to the natural environment.

For these reasons there has been a great deal of interest in studying neural responses to complex, natural stimuli (for example, see refs 1, 21–26). However, the relationship between coding of natural and laboratory stimuli remains elusive due to the difficulty of characterizing neurons—assessing their receptive fields—from responses to natural stimuli, as we now describe.

A simple and commonly used model of neural responses is the linear–nonlinear model^{27,28}. In this model, the response of the neuron depends on linear filtering of the stimulus luminance values $\mathbf{S}$ by a receptive field $\mathbf{L}$ defined over some region of space and time. Mathematically, the filter output at time t is a sum over the spatial positions (x, y) and temporal delays t' to which the neuron's response is sensitive: $\sum_{x, y, t'} \mathbf{L}(x, y, t - t') \mathbf{S}(x, y, t')$, which we abbreviate as $\mathbf{L} * \mathbf{S}$. The output of this filter is then passed through a nonlinear function f to yield the neuron's response r : $r(t) = f(\mathbf{L} * \mathbf{S})$. The nonlinearity incorporates the fact that the firing rate cannot be negative and other aspects of neural response such as threshold, saturation and sensitivity or insensitivity to changes in stimulus polarity. We will use the terms neural filter or receptive field throughout this paper to mean the linear part $\mathbf{L}$ of the linear–nonlinear model.

Traditionally, neural receptive fields have been estimated as the spike-triggered average stimulus (STA; with appropriate correction for autocorrelation of the inputs)^{1,23–25,27,28} or by related methods^{10,26,29}. These methods give unbiased results for linear systems for any stimulus ensemble or for nonlinear systems if the ensemble is gaussian random noise. However, they produce systematic deviations from the true filter of nonlinear 'linear–nonlinear' neurons probed with natural stimuli (or other non-gaussian stimuli), even in situations where the only nonlinearity is due to a conversion of the output of a linear receptive field to firing rate^{24,30}. This happens because natural stimuli, unlike gaussian stimuli which may be completely described by pairwise correlations, have strong higher-order as well as pairwise correlations^{31–33}. The higher-order correlations may be viewed as what distinguishes natural from random gaussian stimuli. The bias in the filter estimate calculated using the gaussian or linear assumption increases with the strength of the nonlinearity and with the strength of stimulus correlations beyond second order^{24,30}, not vanishing even with infinite data.

Recently an information-theoretic method has been developed that correctly estimates receptive fields of nonlinear model neurons (with extensions to multiple linear filters) for arbitrary stimulus ensembles regardless of the strength of multi-point correlations, even in cases where the STA is zero³⁰. According to this method, one searches for the spatiotemporal filter $\mathbf{L}$ whose output, $\mathbf{L} * \mathbf{S}$, carries the most mutual information with the experimentally measured neuronal response $r(t)$. In practice, this is done via a gradient ascent procedure, searching in the space of all possible spatiotemporal receptive fields or filters to find the most informative one (referred to as 'the most informative dimension', or MID). We can then calculate the nonlinearity associated with the MID from the data as the probability of a spike given the filter output; there is no need to make any assumption about the shape of the nonlinearity.

Similarly to other 'spike-triggered' methods, the MID method compares two probability distributions of outputs for a given filter: the distribution of outputs that occur before (or trigger) a spike, and

¹Sloan-Swartz Center for Theoretical Neurobiology and ²Department of Physiology, University of California, San Francisco, 513 Parnassus Avenue, San Francisco, California 94143-0444, USA. ³Center for Theoretical Neuroscience, Center for Neurobiology and Behavior, Columbia University Medical School, N.Y.S.P.I. Kolb Research Annex, 1051 Riverside Drive, Unit 87, New York, New York 10032-2695, USA.

the distribution of outputs over the entire stimulus ensemble regardless of neural response. If a filter represents a stimulus feature that affects neural responses, then certain values of its output will be more probable before a spike, and so the two distributions should differ from one another. The various methods all seek filters that maximize the difference between the two distributions, but differ in the measure of this difference. For the STA, the measure is the change in the mean of the two distributions; for the spike-triggered covariance method^{10,26,29}, it is the change in the variance; and for the MID, it is an information-theoretic measure (the Kullback–Leibler distance) that corresponds to the mutual information between the filter output and the spikes. The information-theoretic measure is more general than the mean or variance, because it is sensitive to correlations of all orders, which in part explains the success of the MID method in estimating neural filters from responses to natural stimuli. Here we apply this method to neural data, focusing on the single-filter model, to address the question of whether and how V1 receptive fields adapt to natural stimuli.

Receptive fields from noise versus natural scenes

We studied 40 simple cells (as characterized by responses to optimal moving gratings³⁴) in anaesthetized cat V1 (complex cells can also be characterized by the MID method³⁰ and will be considered in a future publication). We probed these neurons with natural and white noise inputs. These inputs differ in two important respects. First, they have very different pairwise correlations, which are described by the power

spectra. The power spectrum of a white noise ensemble does not depend on either spatial or temporal frequency within a certain range, whereas the power spectrum of natural inputs depends on spatial frequency k as $\sim 1/k^2$ under a wide variety of conditions^{31,33,35,36} (spatiotemporal statistics have similar structure³⁶). Second, natural scenes have strong statistical correlations beyond second order that cannot be described by the power spectrum, as evident in the much greater incidence of oriented edges in natural scenes than in gaussian noise with the same power spectrum^{31,33}.

To estimate spatiotemporal receptive fields or neural filters from responses to noise and natural stimuli, we applied both the linear systems and information-theoretic methods. The resulting estimated filters and STAs for two example cells are shown in Fig. 1. With respect to responses to the noise ensemble, we found the filter for each cell either as the traditional STA or as the MID³⁰. As expected for white noise stimuli, the two estimates do not differ significantly from each other for the illustrated cells or for most cells ($P > 0.05$ for 31 out of 40 cells, t -test, see Supplementary Methods); the remaining differences can be attributed to the residual spatial correlations in the white noise ensemble. This agreement illustrates the basic validity of the MID method under circumstances where the STA offers an independent unbiased estimate.

For responses to the natural stimulus ensemble, we calculated the STA and corrected it for second-order correlations present in the natural ensemble to obtain a decorrelated STA (dSTA). This would describe the neuron's filter if the neuron were linear. Because this

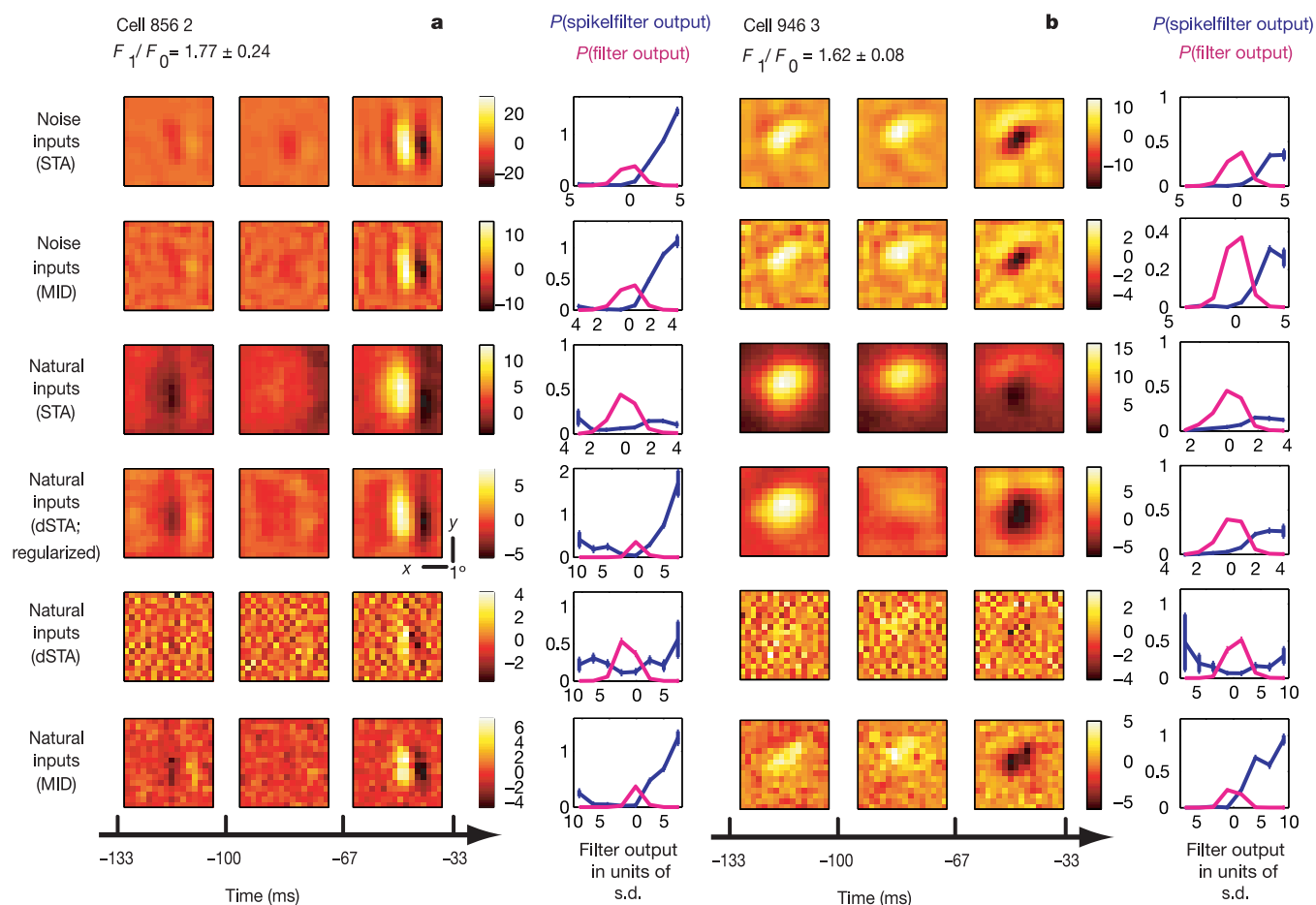


Figure 1 | Filters and nonlinearities for two simple cells. **a, b,** Top to bottom: STA and MID for noise ensemble; STA, dSTA, dSTA with regularization and MID for natural ensemble. Spatiotemporal receptive fields have three time frames covering the indicated time interval (–133 to –33 ms). In the right-most column for each filter we plot the probability

distribution of filter outputs in the stimulus ensemble (magenta) and the spike probability given the filter output (blue; values on the y axis refer to these probabilities). The colour scale shows the filter in units of its average noise level (see Supplementary Methods). x – y scale bars: 1° . Error bars show standard errors of the mean in all figures.

procedure of correcting for stimulus correlations tends to amplify noise, we also calculated the dSTA using regularization to prevent such amplification—such decorrelation with regularization has been used in most previous work estimating neural filters from responses to natural signals^{1,21,24–26}. Finally, we estimated the filter from natural inputs as the MID. As can be seen in Fig. 1, the MID produces an estimate of the filter for natural scenes that is much closer to the white noise filter than either the dSTA or the regularized dSTA. Across cells, the dSTA shows a greater difference from the white noise filter than does the natural ensemble MID, as judged by smaller correlation coefficients with either the noise ensemble STA or noise ensemble MID (40 out of 40 cells, $P < 10^{-6}$). This demonstrates that some of the differences between the neural filters obtained from natural and noise stimulation in the linear model are due to biases in the estimation of the natural filter that can be removed once the linear–nonlinear model is considered and the MID is computed. In Fig. 1, we also plot the nonlinear functions that show spike probability as a function of filter output. They are similar in shape for the MIDs of the two ensembles, and this behaviour seems to be typical across cells.

We used the MIDs to estimate both the noise and natural filters in what follows. We studied all simple cells with a non-zero filter to both natural and noise inputs.

Despite the similarity of the filters obtained under the two conditions (see Fig. 1), a jackknife analysis of the errors in estimating the neural filters shows that the differences between the filters derived from noise and natural signals are statistically significant ($P < 0.01$) for all cells. To investigate the source of these differences and to make connections with classical studies on neural responses to moving periodic patterns (gratings) of certain orientations and spatial frequencies, we compute the spatiotemporal Fourier transform of the filter in the two spatial dimensions and time. The position of the maximum of the Fourier transform at the grating temporal frequency is our prediction for the optimal grating orientation and

spatial frequency for a particular neuron. We did not detect any systematic shifts in optimal orientation and only a small shift in optimal spatial frequency as assayed from noise filters, natural signal filters and grating stimuli, in agreement with previous findings using the regularized dSTA^{25,26} (see Supplementary Discussion).

The most marked differences between the neural filters derived from natural versus noise stimulation are seen by considering the entire shape of the spatial frequency tuning curves (Fig. 2 and Supplementary Figs 1 and 2) and not just the location of the single best spatial frequency. For each cell and temporal frequency, we calculated the spatial frequency profile along the cell's preferred stimulus orientation using interpolation of the filter's two-dimensional discrete Fourier transform. Note that our temporal resolution allowed analysis only at two temporal frequencies (0 Hz and 10 Hz in each of two opposite directions of motion). Results at 10 Hz did not depend on direction of motion, so both directions were combined in Fig. 2, which shows the average tuning of the cells in our data set. For low spatial frequencies sensitivity decreased (increased) to common (rare) inputs, whereas at middle and high spatial frequencies the sensitivity did not change. For example, at zero temporal frequency, low spatial frequencies are more common in the natural than in the white noise stimulus ensemble (Fig. 2b). Correspondingly, neurons became less sensitive to those frequencies during stimulation with natural inputs than during stimulation with noise inputs (Fig. 2a). In the case of non-zero temporal frequencies the trend is reversed, because the noise stimulus ensemble has more power at nearly all spatial frequencies than the natural stimulus ensemble (Fig. 2d, e). These changes in filter can be observed in the majority of cells, and are not simply due to adaptation in a small subset of cells. This is shown in Supplementary Fig. 1, which illustrates the spatial frequency sensitivities of the two example cells for which the receptive fields are shown in Fig. 1, and Supplementary Fig. 2, which shows scatter plots of spatial frequency sensitivity of noise versus natural filters across all cells.

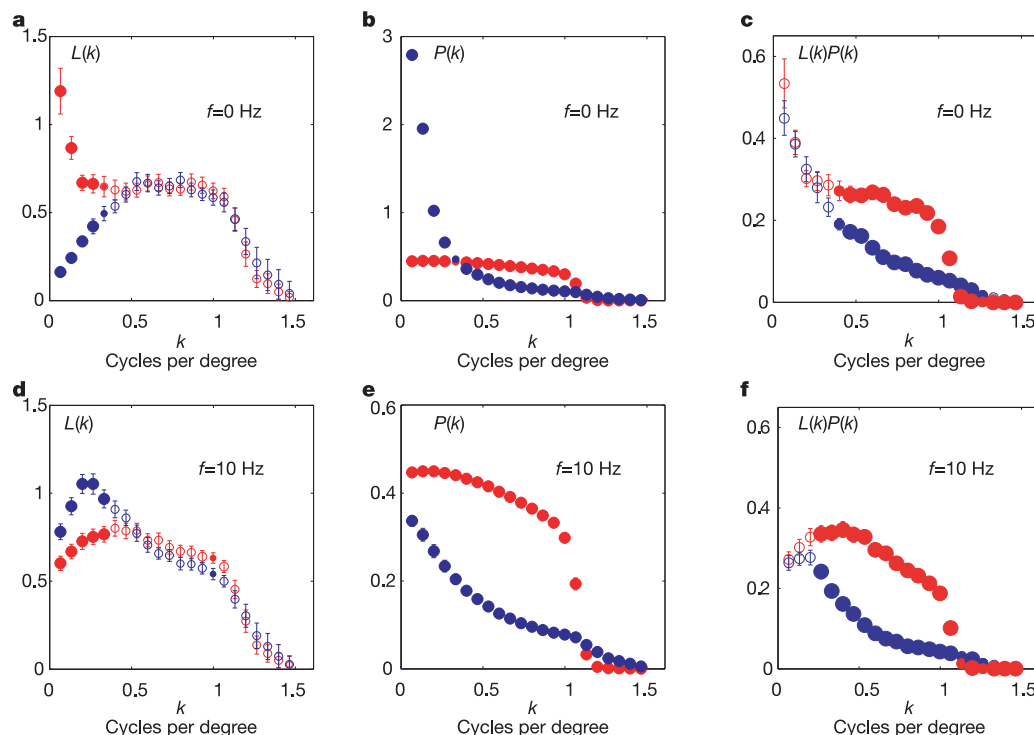


Figure 2 | Neural filters compensate for changes in the input power spectrum. Average amplitude spectra of neural filters (**a, d**) and input ensembles (**b, e**) corresponding to natural (blue circles) and white noise (red circles) stimulation for temporal frequencies of 0 and 10 Hz. The spectra were taken along the optimal orientation for each cell by interpolating the

discrete two-dimensional Fourier transform. We use filled circles at frequencies where mean sensitivity was significantly different between the two ensembles (small circles for $P < 0.05$ and large circles for $P < 0.01$), and open circles otherwise. **c, f**, Plots of the product of the average neural filter and input ensemble amplitude spectra.

Optimal filtering in a nonlinear system

In retrospect, such shifts in spatial frequency sensitivity may be expected for neural coding to be optimal for both of two input ensembles (white noise and natural stimuli) that have such vastly different power spectra^{31,33,35} (compare Fig. 2b, e). In general it is difficult to map optimal coding strategy from one ensemble to another; however, it could be done if both of the stimulus ensembles were gaussian so that they were entirely characterized by their power spectra. Suppose a neuron uses filter L_A and nonlinearity f_A to encode optimally gaussian stimulus ensemble A with spatiotemporal amplitude spectrum $P_A(k, \omega)$. What would then be an optimal strategy to encode gaussian ensemble B with amplitude spectrum $P_B(k, \omega)$? One solution is to leave the nonlinearity unchanged and to compensate for differences in the input power spectra by changing neural filter properties so that:

$$L_A(k, \omega)P_A(k, \omega) = L_B(k, \omega)P_B(k, \omega) \quad (1)$$

This will leave unchanged all statistics of neuronal response, and so in particular will leave invariant any statistical measures of optimality. Alternative strategies involving a change in nonlinearity cannot be optimal unless there are multiple optima, because if ensemble A has a unique optimum, then the above strategy will give the unique optimum for ensemble B. (Note that, in response to an overall change in contrast, the nonlinearity can be rescaled^{10,12}, but this is equivalent to a rescaling of the filter according to equation (1) with no change in nonlinearity.)

These conclusions about the receptive field and nonlinearity apply only to gaussian stimuli. The higher-order correlations present in natural scenes may both lead to deviations from equation (1) in neural filters and cause changes in the shape of the nonlinearity. But in practice, the changes in the shape of the nonlinearity are small, and changes in neural filters that do take place act to compensate for changes in the input power spectrum as predicted from equation (1) (Fig. 2c, f). These changes in frequency sensitivity occur primarily at low spatial frequencies. No changes are observed at mid-to-high spatial frequencies, resulting in significant deviations from equation (1) in the middle range of frequencies. We can only speculate that other factors may limit the range of spatial frequencies over which adaptation can occur.

Adaptation increases information transmission

The above optimal coding argument provides at least a qualitative explanation of observed receptive field changes. Most theories of optimal coding define optimality in information-theoretic terms. To test directly whether the information maximization argument applies to our data, we calculated the average mutual information

between the filter output and the neural response; the response at a given time is simply taken as the presence or absence of a single spike³⁰.

The changes in receptive fields act to increase the information after changes in stimulus ensemble, and this information would be substantially reduced if receptive fields did not change with the ensemble. That is, the natural filter carries more information about responses to the natural ensemble than to the noise ensemble ($P < 10^{-4}$, paired Wilcoxon two-tailed test), whereas the noise filter carries more information about responses to the noise ensemble than to the natural ensemble ($P = 0.03$). The average information values across the population are shown in Fig. 3, and scatter plots on a cell-by-cell basis are provided in Supplementary Fig. 4. Each filter produces roughly equal information about responses to its own ensemble: the difference in information values achieved by applying the noise filter to the noise ensemble versus applying the natural filter to the natural ensemble is not significant ($P = 0.18$, paired Wilcoxon test). Each filter produces substantially less information about responses to the other ensemble ($P < 10^{-4}$ for natural or noise ensemble filtered with natural versus noise filter; paired Wilcoxon tests), and there is no significant difference between the swapped combinations (natural filter applied to noise ensemble or vice versa, $P = 0.06$, paired Wilcoxon test). We note that the changes in information are not due to overfitting or other computational artefacts, because information was calculated from responses to ensemble segments that were not used in calculating the filters, and the effects were not seen in data from a model linear–nonlinear cell with unchanging filter that was analysed similarly (see Supplementary Information).

In addition to considering information I in bits (Fig. 3, top), we also measured information for each cell in units of I_{spike} ³⁷, the information in the neuron's response (as defined above) about the full, unfiltered stimulus (Fig. 3, bottom). I/I_{spike} measures the fraction of the total possible information that is captured by the single most informative filter (I_{spike} is a separate measurement that was available only for a subset of cells, making the data set smaller). As can be seen, the MID captures roughly 35% of the possible information for simple cells. Each filter provides a greater fraction of the overall information when applied to its own ensemble than the other ($P \leq 10^{-4}$ for natural filter applied to natural versus noise ensemble and for either ensemble filtered with natural versus noise filter; $P = 0.05$ for noise filter applied to natural versus noise ensemble; paired Wilcoxon test).

Dynamics of receptive field adaptation

Even though the best linear–nonlinear model systematically changes with the stimulus ensemble, this does not establish that the neuron has changed its encoding strategy. The true encoding strategy may be complicated and nonlinear, so that even if it is static, the best linear–nonlinear estimate of it may change with the ensemble, much as the best linear approximation to a curve changes with position on the curve.

The most direct method to distinguish between an adaptive strategy and a complex but static coding strategy would be to estimate the filter as a function of time and see it change. This method yields very poor time resolution, because ~ 5 min of data are needed to estimate the filter, so adaptation that occurs on a faster timescale cannot be seen. Nonetheless we tried this method and saw appropriate, if weak, adaptation to noise stimuli even on this long timescale (see Supplementary Fig. 5). To achieve finer time resolution, we studied adaptation by measuring changes with time in the information carried by the output of a single, static filter; this information can be estimated from ~ 30 s of data. We used the following reasoning. If the coding is static, then the mutual information between this filter's output and the neuron's responses to a given ensemble should not systematically change in time. However, if the neuron's receptive field adapts to the stimulus

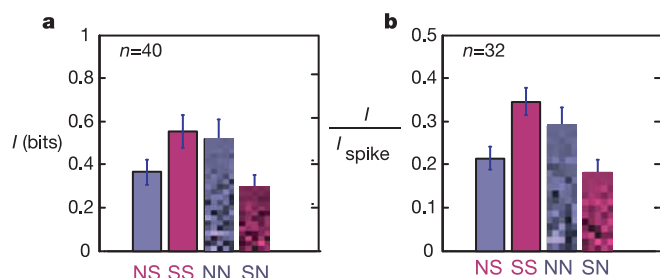


Figure 3 | Receptive field adaptation increases information transmission. Bars show the mutual information between spikes and outputs of either the noise (blue, N) or natural scenes (red, S) filter applied to the natural scenes ensemble (solid) or noise ensemble (pixelated). NS, white noise filter applied to natural scenes ensemble; SS, natural scenes filter applied to natural scenes ensemble; NN, white noise filter applied to noise ensemble; SN, natural scenes filter applied to noise ensemble. The information values are given in bits (a) or in units of the total information carried by the arrival of a single spike I_{spike} ³⁷ (b).

ensemble, then this information may systematically change in time. In particular, we take the static filter to be that characterizing a neuron when it is well adapted to a given ensemble—say the natural ensemble. When the neuron is newly exposed to a natural ensemble, the information carried by this filter should increase with increasing time of exposure to the natural ensemble, as the neuron adapts so that the filter that it actually uses to encode incoming stimuli into spikes becomes closer and closer to this static, fully adapted filter. Similarly, when the neuron is newly exposed to a noise ensemble, the information carried by this filter should decrease with increasing time of exposure to the noise ensemble, as the neuron's own filter adapts to the noise and becomes less and less like the fully adapted natural scenes filter.

We derived filters from the last half of the 10-min presentations of each stimulus ensemble, when the neuron would be best adapted to the given ensemble if adaptation occurs. We then applied these static filters to both noise and natural stimuli, and measured information between spikes and filtered stimuli in successive 34-s periods during the first half of stimulus presentation (if the filter was derived from the second half of this stimulus) or in successive 68-s periods during all of the presentation of the opposite ensemble. Most cells did not show significant adaptation when considered individually, presumably due to the variability in measuring information over such brief time periods. However, averaging over the entire population of simple cells revealed clear adaptation over time, consistent with an adaptive coding strategy (Fig. 4). The information progressively increased with time when natural inputs were filtered with the neural filter derived from the natural stimulus ensemble (Fig. 4a; see also Supplementary Discussion and Supplementary Fig. 6), whereas the information decreased with time when that same filter was applied to noise inputs (Fig. 4b).

Fits of a single exponential to the average data demonstrate that there is a statistically significant monotonic change with time, with time constants $\tau = 42 \pm 9$ s for adaptation to the natural ensemble and $\tau = 22 \pm 2$ min for adaptation to the noise ensemble. These time constants are consistent with the fact that we could not detect adaptation to the natural ensemble with the 5-min timescale of direct filter measurements, but we could detect adaptation to the noise ensemble (Supplementary Fig. 5). Note, however, that the time

constants are based on the assumption of exponential decay and do not exclude the possibility of multiple timescales, including scales faster than we were able to measure, or of alternative functional forms of decay.

We could not detect a significant trend with time in the information carried by the noise filter about either ensemble (see Supplementary Fig. 7). This is perhaps not surprising given that the average decrease in information for the noise filter applied to the noise versus natural ensembles was not significant ($P = 0.14$, unpaired t -test), and that the slow time course of adaptation to the noise ensemble suggests that the filter we tested was not fully adapted to it (see also the legend of Supplementary Fig. 7). Nonetheless, the presence of significant monotonic changes in the expected directions for the natural scenes filter applied to each ensemble demonstrates that the neuron's coding strategy is adapting over time with exposure to a given ensemble.

Discussion

Adaptation is ubiquitous throughout the nervous system, and it occurs in many forms. In vision, adaptation to luminance mean and variance (contrast) has been observed in the retina^{5,6,9,13,14,17,18}, lateral geniculate nucleus¹⁶ and primary visual cortex^{4,7,8,20}, and related changes are observed in perception³⁸. In the framework of our model, adaptation may affect the neural gain (the nonlinear input–output function) or the spatiotemporal filter itself. Adaptation of the gain to the mean and variance of the stimulus ensemble (and perhaps to higher-order statistics³⁹) serves to fit a neuron's dynamic range to the dynamic range of the stimulus^{5–7,9,10,12–14,16}. In addition, adaptation of the filter to the mean and variance of the stimulus^{2,3,5,6,14,17} has been observed, and it has been argued that such adaptation along with adaptation to the stimulus covariance can serve to maximize the information per spike in the neuron's response^{40,41}. In general, filter adaptations are nearly instantaneous (<0.1 s), whereas changes in gain can be more gradual (time constants up to 10 s and perhaps longer for some components of adaptation to mean luminance)^{5,6,9,13,14,16,17}. Here we find an adaptive change in neural filters in response to stimulus statistics beyond the mean and variance, and one that occurs over much longer timescales than previously found even for contrast gain changes. This suggests that

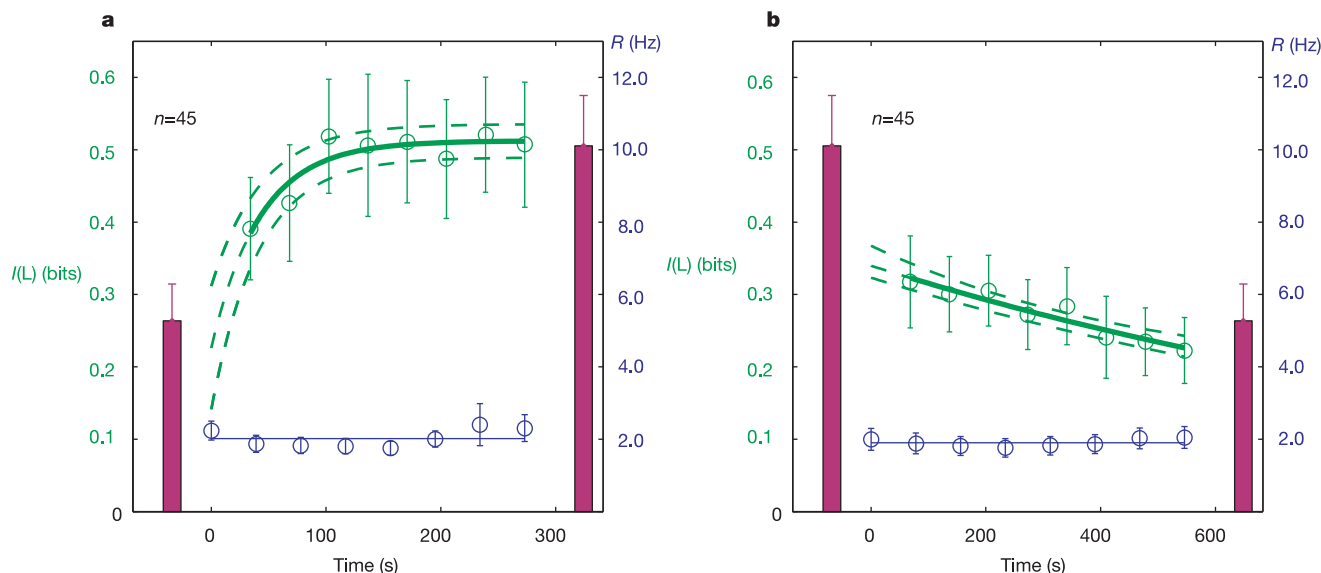


Figure 4 | Adaptation dynamics. **a, b**, The neural filter derived from the last half of natural stimulation is applied to the first half of natural (**a**) or to the noise (**b**) ensemble. Symbols show information (green, left y axis) and firing rate (blue, right y axis) averaged across cells, versus time. The solid line is an exponential fit; dashed lines show one standard deviation based on the

Jacobian of the fit ($P = 0.01$ in **a** and $P = 0.003$ in **b** using an F -test with null hypothesis of no time dependence). The taller (shorter) red bars show information for the natural filter applied to natural (noise) inputs (as in Fig. 3, but $n = 45$). The firing rates demonstrate that recordings are stable.

the observed adaptation represents a new mechanism for optimal coding.

Adaptation to the power spectrum could be considered a generalized form of contrast adaptation, in which different frequency channels providing input to cortical cells differentially adapt their gains so that channels with more stimulus power show greater adaptation. Indeed, variation of gain adaptation across different retinal pathways has been observed^{9,13,16,18}. However, these observations, and a recently reported pattern-specific component of retinal adaptation⁴², involved adaptation on significantly faster time-scales than observed here. Also, in the lateral geniculate nucleus, adaptive changes between white noise and natural stimulation were not observed in the temporal domain, at least for a majority of cells⁴³. This suggests that the adaptive changes reported here are of cortical origin. A pattern-specific component of cortical adaptation has been observed: for example, one that differentially affects responses according to the difference of the stimulus orientation, direction or spatial frequency from that of the adapting stimulus^{8,11,15,19,20}. At least in one case, this adaptation has been observed to have time constants on the order of a minute or longer¹¹. It is possible that the present observations may share some underlying mechanisms with such pattern-specific adaptation.

Many recent studies have used versions of the linear model or related models to estimate receptive fields from responses to natural stimuli^{1,21,23–26}. Some have reported that the estimates calculated from responses to natural stimuli differ from those calculated from responses to noise^{1,21,23}, whereas others^{25,26} found no change in the major parameters of neural filters, such as optimal stimulus orientation and spatial frequency. It is not clear from these observations to what degree reported differences in neural filters are genuinely stimulus-induced or are due to biases in the estimation induced by the non-gaussian statistics of natural stimuli together with the nonlinearity of the input–output function. The fact that the receptive field obtained for a given ensemble from the linear model best predicted responses to other examples of its own ensemble^{1,21,23} suggests at least partially genuine differences, which is also supported by our results on spatial frequency adaptation. However, the fact that we found larger differences between filters obtained in the linear approximation (dSTA for natural stimulus ensemble and STA for white noise ensemble) than between filters obtained in the linear–nonlinear model (MID for natural stimulus ensembles and STA or MID for white noise ensemble) suggests that biases also exist, and the new information maximization procedure used here removes these biases for real neurons, just as was demonstrated in numerical simulations³⁰.

We have found that V1 neurons adapt their filters to stimulus statistics beyond the mean and variance. This filter adaptation occurs over 40 s to many minutes, suggesting that it is not a consequence of previously described mechanisms of luminance or contrast adaptation. The adaptation serves to preserve information transmission and to reduce relative responses to stimulus components that are relatively more abundant in the stimulus ensemble, as predicted by optimal encoding arguments. It remains to be determined whether the neurons are adapting to changes in power spectra, in higher-order statistics, or both. The gradual nature of adaptive changes and their correspondence to optimization principles suggests that it might be possible to predict the direction and degree of adaptation to stimulus sets with statistics intermediate between those of white noise and natural stimuli. Thus, there is hope for creating a unified picture of neural responses across various input ensembles.

METHODS

All experimental recordings were conducted under a protocol approved by the University of California, San Francisco, Committee on Animal Research with procedures previously described⁴⁴. Spike trains were recorded using tetrode electrodes from the primary visual cortex of anaesthetized adult cats and

manually sorted off-line. Visual stimulus ensembles of white noise and natural scenes were each 546 s long. After manually estimating the size and position of the receptive field, neurons were probed with full-field moving periodic patterns (gratings). Cells were selected as simple if, under stimulation by a moving sinusoidal grating with optimal parameters, the ratio of their response modulation (F_1 ; that is, amplitude of the Fourier transform of the response at the temporal frequency of the grating) to the mean response (F_0) was larger than one³⁴. The rest of the protocol typically consisted of an interlaced sequence consisting of three different noise input ensembles of identical statistical properties, and three different natural input ensembles. The interval between presentations varied in duration as necessary to provide adequate animal care. All natural input ensembles were recorded in a wooded environment with a hand-held digital video camera in similar conditions on the same day (see Supplementary Movie). The noise ensembles were white overall, but the spatial frequency spectrum was divided into eight circular bands, and each particular frame was limited to one band chosen at random; this white noise design was intended to increase the number of elicited spikes. The mean luminance and contrast of the noise ensembles were adjusted to match those of the natural ensembles. Both noise and natural inputs were shown at 128×128 pixel resolution, with angular resolution of approximately 0.12° per pixel. To calculate receptive fields, input ensembles were down-sampled to 32×32 pixels. The receptive field centre was determined from the maxima in the STAs for noise and natural ensembles and was set to the same position for analysis of both noise and natural inputs. A patch of 16×16 pixels was selected around the centre (angular resolution of 0.48° per pixel) to make analysis computationally feasible and to minimize effects due to undersampling (we strove to have the number of spikes greater than the dimensionality of the receptive fields³⁰). In all cases subsequent analysis of receptive fields verified that the selected patch fully contained the receptive field. These receptive fields were used in all quantitative analyses (Figs 2–4). Examples in Fig. 1 were computed at and are shown at twice the angular resolution to illustrate the finer structure of the receptive field, as well as differences in performance of the various methods.

Received 27 April; accepted 1 December 2005.

1. Theunissen, F. E., Sen, K. & Doupe, A. J. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.* **20**, 2315–2331 (2000).
2. Sceniak, M. P., Hawken, M. J. & Shapley, R. Contrast-dependent changes in spatial frequency tuning of macaque V1 neurons: effects of a changing receptive field size. *J. Neurophysiol.* **88**, 1363–1373 (2002).
3. Nolt, M. J., Kumbhani, R. D. & Palmer, L. A. Contrast-dependent spatial summation in the lateral geniculate nucleus and retina of the cat. *J. Neurophysiol.* **92**, 1708–1717 (2004).
4. Maffei, L., Fiorentini, A. & Bisti, S. Neural correlate of perceptual adaptation to gratings. *Science* **182**, 1036–1038 (1973).
5. Shapley, R. & Victor, J. D. The contrast gain control of the cat retina. *Vision Res.* **19**, 431–434 (1979).
6. Shapley, R. M. & Enroth-Cugell, C. in *Progress in Retinal Research* (eds Osborne, N. N. & Chader, G. J.) 263–346 (Pergamon, Oxford, 1984).
7. Ohzawa, I., Sclar, G. & Freeman, R. D. Contrast gain control in the cat's visual system. *J. Neurophysiol.* **54**, 651–667 (1985).
8. Saul, A. B. & Cynader, M. S. Adaptation in single units in visual cortex: The tuning of aftereffects in the spatial domain. *Vis. Neurosci.* **2**, 593–607 (1989).
9. Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W. & Meister, M. Adaptation of retinal processing to image contrast and spatial scale. *Nature* **386**, 69–73 (1997).
10. Brenner, N., Bialek, W. & de Ruyter van Steveninck, R. Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702 (2000).
11. Dragoi, V., Sharma, J. & Sur, M. Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron* **28**, 287–298 (2000).
12. Fairhall, A. L., Lewen, G. D., Bialek, W. & de Ruyter Van Steveninck, R. R. Efficiency and ambiguity in an adaptive neural code. *Nature* **412**, 787–792 (2001).
13. Chander, D. & Chichilnisky, E. J. Adaptation to temporal contrast in primate and salamander retina. *J. Neurosci.* **21**, 9904–9916 (2001).
14. Baccus, S. A. & Meister, M. Fast and slow contrast adaptation in retinal circuitry. *Neuron* **36**, 909–919 (2002).
15. Kohn, A. & Movshon, J. A. Adaptation changes the direction tuning of macaque MT neurons. *Nature Neurosci.* **7**, 764–772 (2004).
16. Solomon, S. G., Peirce, J. W., Dhruv, N. T. & Lennie, P. Profound contrast adaptation early in the visual pathway. *Neuron* **42**, 155–162 (2004).
17. Victor, J. D. The dynamics of the cat retinal x-cell centre. *J. Physiol. (Lond.)* **386**, 219–246 (1987).
18. Brown, S. P. & Masland, R. H. Spatial scale and cellular substrate of contrast adaptation by retinal ganglion cells. *Nature Neurosci.* **4**, 44–51 (2001).
19. Movshon, J. A. & Lennie, P. Pattern-selective adaptation in visual cortical neurones. *Nature* **278**, 850–852 (1979).
20. Albrecht, D. G., Farrar, S. B. & Hamilton, D. B. Spatial contrast adaptation

- characteristics of neurones recorded in the cat's visual cortex. *J. Physiol. (Lond.)* **347**, 713–739 (1984).
21. David, S. V., Vinje, W. E. & Gallant, J. L. Natural stimulus statistics alter the receptive field structure of V1 neurons. *J. Neurosci.* **24**, 6991–7006 (2004).
 22. Baddeley, R. *et al.* Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lond. B* **264**, 1775–1783 (1997).
 23. Theunissen, F. E. *et al.* Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network* **12**, 289–316 (2001).
 24. Ringach, D. L., Hawken, M. J. & Shapley, R. Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *J. Vis.* **2**, 12–24 (2002).
 25. Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D. & Tolhurst, D. J. The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *J. Neurosci.* **23**, 4746–4759 (2003).
 26. Felsen, G., Touryan, J., Han, F. & Dan, Y. Cortical sensitivity to visual features in natural scenes. *PLoS Biol.* **3**, e342 (2005).
 27. de Boer, R. & Kuyper, P. Triggered correlation. *IEEE Trans. Biomed. Eng.* **15**, 169–179 (1968).
 28. Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. B. *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, Massachusetts, 1997).
 29. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque v1 receptive fields. *Neuron* **46**, 945–956 (2005).
 30. Sharpee, T., Rust, N. C. & Bialek, W. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.* **16**, 223–250 (2004).
 31. Ruderman, D. L. & Bialek, W. Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.* **73**, 814–817 (1994).
 32. Field, D. J. What is the goal of sensory coding? *Neural Comput.* **6**, 559–601 (1994).
 33. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
 34. Skottun, B. C. *et al.* Classifying simple and complex cells on the basis of response modulation. *Vision Res.* **31**, 1079–1086 (1991).
 35. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
 36. Dong, D. W. & Atick, J. J. Statistics of natural time-varying images. *Network* **6**, 345–358 (1995).
 37. Brenner, N., Strong, S. P., Koberle, R., Bialek, W. & de Ruyter van Steveninck, R. R. Synergy in a neural code. *Neural Comput.* **12**, 1531–1552 (2000).
 38. Blakemore, C. & Campbell, F. W. Adaptation to spatial stimuli. *J. Physiol. (Lond.)* **200**, 11P–13P (1969).
 39. Kvale, M. N. & Schreiner, C. E. Short-term adaptation of auditory receptive fields to dynamic stimuli. *J. Neurophysiol.* **91**, 604–612 (2004).
 40. Wainwright, M. J. Visual adaptation as optimal information transmission. *Vision Res.* **39**, 3960–3974 (1999).
 41. Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).
 42. Hosoya, T., Baccus, S. A. & Meister, M. Dynamic predictive coding by the retina. *Nature* **436**, 71–77 (2005).
 43. Dan, Y., Atick, J. J. & Reid, R. C. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* **16**, 3351–3362 (1996).
 44. Emondi, A. A., Rebrik, S. P., Kurgansky, A. V. & Miller, K. D. Tracking neurons recorded from tetrodes across time. *J. Neurosci. Methods* **135**, 95–105 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge suggestions from W. Bialek on the design of experiments and subsequent data analysis. We thank M. Caywood, B. St Amant and K. MacLeod for help with experiments. We thank P. Sabes, M. Kvale and S. Palmer for helpful suggestions on statistical aspects of data analysis. Computing resources were provided by the National Science Foundation under the following NSF programs: Partnerships for Advanced Computational Infrastructure at the San Diego Supercomputer Center through an NSF cooperative agreement, Distributed Terascale Facility (DTF) and Terascale Extensions (enhancements to the Extensible Terascale Facility). This research was supported through a grant to K.M. from the National Eye Institute and by a grant from the Swartz Foundation and a career development award from the National Institute of Mental Health to T.S.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.S. (sharp@phy.ucsf.edu).

Discovery of two new satellites of Pluto

H. A. Weaver¹, S. A. Stern², M. J. Mutchler³, A. J. Steffl², M. W. Buie⁴, W. J. Merline², J. R. Spencer², E. F. Young² & L. A. Young²

Pluto's first known satellite, Charon, was discovered¹ in 1978. It has a diameter ($\sim 1,200$ km) about half that of Pluto^{2–4,17}, which makes it larger, relative to its primary, than any other moon in the Solar System. Previous searches for other satellites around Pluto have been unsuccessful^{5–7}, but they were not sensitive to objects $\lesssim 150$ km in diameter and there are no fundamental reasons why Pluto should not have more satellites⁶. Here we report the discovery of two additional moons around Pluto, provisionally designated S/2005 P 1 (hereafter P1) and S/2005 P 2 (hereafter P2), which makes Pluto the first Kuiper belt object known to have multiple satellites. These new satellites are much smaller than Charon, with estimates of P1's diameter ranging from 60 km to 165 km, depending on the surface reflectivity; P2 is about 20 per cent smaller than P1. Although definitive orbits cannot be derived, both new satellites appear to be moving in circular orbits in the same orbital plane as Charon, with orbital periods of ~ 38 days (P1) and ~ 25 days (P2).

We observed Pluto with the Hubble Space Telescope using the Wide-Field Channel (WFC) mode of the Advanced Camera for Surveys (ACS) on UT 2005 May 15 and May 18 (Fig. 1). The ACS/WFC consists of two $4,096 \times 2,048$ pixel CCDs (WFC1 and WFC2) butted together, effectively forming a single $4,096 \times 4,096$ pixel camera with a gap of ~ 50 pixels between the two CCDs. The F606W ('Broad V') filter, which has a centre wavelength of 591.8 nm and a width of 67.2 nm, was used for all images. At the time of the observations, Pluto was 31.0 astronomical units (AU) from the Sun, 30.1 AU from the Earth, and had a solar phase angle of 0.96° on 15 May and 0.88° on 18 May. Identical strategies were employed on each observing date. First, a single short exposure (0.5 s) was taken to enable accurate positions of Pluto and Charon to be measured on unsaturated images. Then, two identical, long exposures (475 s) were taken at the same pointing to provide high sensitivity to faint objects. Finally, the telescope was moved by ~ 5 pixels in one dimension and ~ 60 pixels in the other dimension, and two identical, long exposures

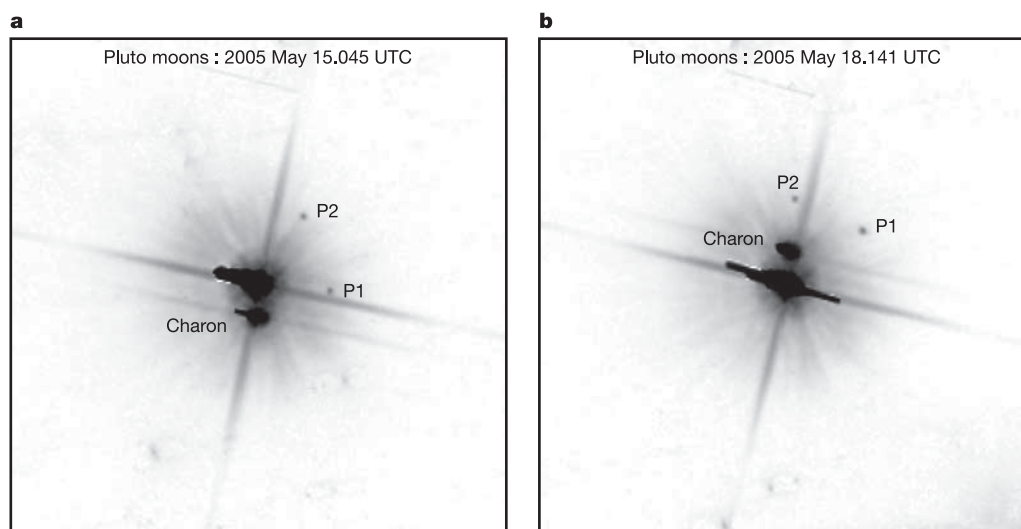


Figure 1 | Images of the Pluto–Charon system, showing the new satellites, S/2005 P 1 (P1) and S/2005 P 2 (P2). **a**, Data taken on 2005 May 15.045 UTC; **b**, Data taken 3.1 days later on 2005 May 18.141 UTC. Each frame is a small section (256 pixel $\times$ 256 pixel) of the full image that is centred on Pluto and subtends 12.8 arcsec $\times$ 12.8 arcsec, which projects to $280,000$ km $\times$ $280,000$ km at Pluto. Celestial north is up and east is to the left. A logarithmic intensity stretch is used to enhance the visibility of the two new satellites in the presence of the much brighter, saturated images of Pluto and Charon. P1 is near the 3 o'clock position on 15 May and moves to the 2 o'clock position 3.1 days later, while P2 moves from the 1 o'clock position to the 12 o'clock position over the same time interval. Charon has moved (also anticlockwise) from one side of Pluto to the other between the two epochs, as expected, owing to its orbital period of 6.4 days. We used the individual, bias-subtracted,

flat-fielded files from the ACS calibration pipeline as the starting point for our data reduction. We coaligned the four long-exposure images taken on each date and combined them using a sigma-clipping procedure that filters out anomalous features not present (to within the noise level) in all images. Thus, the many camera artefacts present in the flat-fielded images (cosmic ray events, hot pixels, dead pixels, and so on), as well as most of the star trails, are removed in the composite image. The few remaining faint blotches are artefacts caused by incomplete subtraction of star-trails in the field, and their spatial distributions are very different from the PSF-like distributions of P1 and P2. The geometric distortion in the composite image was removed using the same remapping algorithm employed by the standard ACS/WFC calibration process.

¹The Johns Hopkins University Applied Physics Laboratory, Space Department, 11100 Johns Hopkins Road, Laurel, Maryland 20723-6099, USA. ²Department of Space Studies, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, Colorado 80302, USA. ³Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ⁴Lowell Observatory, 1400 W. Mars Hill Road, Flagstaff, Arizona 86001, USA.

Table 1 | Astrometry and photometry of Pluto satellites

Date (UT May 2005)	Pluto		Charon		S/2005 P1		S/2005 P2	
	[<i>r</i> , <i>θ</i>]	<i>V</i>	[<i>r</i> , <i>θ</i>]	<i>V</i>	[<i>r</i> , <i>θ</i>]	<i>V</i>	[<i>r</i> , <i>θ</i>]	<i>V</i>
15.045	[0, 0]	14.12 ± 0.08	[0.834, 176.77]	16.22 ± 0.10	[1.85, 264.21]	NA	[2.09, 326.94]	23.38 ± 0.17
18.141	[0, 0]	14.28 ± 0.08	[0.855, 355.04]	16.28 ± 0.10	[2.36, 305.76]	22.93 ± 0.12	[2.22, 355.51]	NA

'Date' refers to the mid-point of the observing period. *r* is the distance in arcsec, and *θ* is the celestial position angle in degrees (measured eastward from north) in the J2000 reference frame from the centre-of-light of the specified object to the centre-of-light of Pluto. The 1σ errors in *r* and *θ* are 0.035 arcsec and 0.7°, respectively, and are mainly due to the uncertainty in locating Pluto in the long-duration exposures. The centre-of-light is offset from the centre-of-body by ~100 km for Pluto and by ~20 km for Charon², but these offsets were ignored in our orbital analyses. *V* is the observed visual magnitude of the object, when available, and was calculated using a 5-pixel-radius (0.25 arcsec) aperture. We used the information and procedures described in ref. 15 to convert the observed signal in this small aperture to a *V*-magnitude in the standard Johnson-Landolt system. Adopting a *B* – *V* colour of 0.7 (that is, similar to Charon-like¹² and solar-like¹⁶ colours), we derive $V = -2.5 \log S/t + (26.38 \pm 0.07)$, where *V* is the *V*-band magnitude in the Johnson-Landolt system, *S* is the net signal in electrons in the 5-pixel-radius aperture, and *t* is the exposure time in seconds. We calculated the magnitudes of Pluto and Charon in the same way, although both objects (especially Pluto) are slightly resolved, which introduces an additional systematic error in those cases. NA, not available.

(475 s) were taken to provide data in the region of the sky falling in the inter-chip gap during the first two long exposures. The telescope was programmed to track the apparent motion of Pluto (~3 arcsec h⁻¹) for all exposures.

The two new satellites are detected with high signal-to-noise ratio (*S/N* ≥ 35) and have a spatial morphology consistent with the ACS point spread function (PSF; this is the spatial brightness distribution expected for unresolved objects). Unlike virtually all of the spurious features in the individual ACS images, the two new objects were observed in two consecutive images on two different dates. Furthermore, the objects do not appear in the images where they are not expected, namely when the telescope pointing placed the objects in the inter-chip gap. Although P1 and P2 are several thousand times fainter than Pluto, their images look nothing like the well-documented⁸ ghosts and scattered-light artefacts produced near bright, highly saturated objects. In short, all the available evidence supports the claim that the objects observed near Pluto–Charon are real astronomical objects, and none of the data suggest that these objects are observational artefacts.

Next we address the issue of whether P1 and P2 are associated with the Pluto–Charon system. Only astronomical bodies moving at approximately the same non-sidereal rate as Pluto will have PSF-like spatial distributions, and asteroids and stars are easily recognized by their trailed images. We estimate that foreground or background objects must have a heliocentric distance within ~0.25 AU of Pluto's to remain within 2 arcsec of Pluto over the 3.1 days between the two sets of ACS observations. On the basis of the known sky-projected density of Kuiper belt objects (KBOs) as a function of magnitude⁹, the probability of finding any similarly bright KBO within 2.5 arcsec of Pluto is ~4 × 10⁻⁶, which means that the probability of finding two KBOs in this same region is ~1.6 × 10⁻¹¹. These statistical arguments are not strictly applicable to the plutino KBO population (that is, objects that share Pluto's 3:2 orbital resonance with Neptune), for which resonances and other gravitational perturbations might keep objects relatively close to Pluto. Nevertheless, there must be only a vanishingly small chance that the two observed objects are random KBOs that happen to be aligned with Pluto during the course of our observations.

Orbital analysis based on the astrometry of P1 and P2 (Table 1) provides even stronger evidence that the two objects are indeed associated with Pluto. Although the astrometric data from two observations alone are not enough to compute definitive orbits, we investigated the hypothesis that the objects have circular orbits in the same plane as Charon's orbit. Adding these two constraints (that is, zero eccentricity and the same orbital plane as Charon) allows us to solve for the orbital radii that best match the observations. We find that the orbital radii are 64,700 ± 850 km for P1 and 49,400 ± 600 km for P2, both measured relative to the system barycentre. The results are depicted in Fig. 2. The excellent agreement between the observed and model orbits (reduced $\chi^2_\nu \approx 1$ for the fits) supports our assumption that the orbits are circular and coplanar with Charon's orbit, and also reinforces the conclusion that P1 and

P2 are associated with the Pluto system. Using the known mass of the Pluto system and Kepler's third law, we derive orbital periods of 38.2 ± 0.8 days for P1 and 25.5 ± 0.5 days for P2. The ratios of the orbital periods to the orbital period of Charon (6.387245 ± 0.000012 days)² are 5.98 ± 0.12 for P1 and 3.99 ± 0.07 for P2. This suggests that P1 may be in a 6:1, and P2 may be in a 4:1, mean motion orbital resonance with Charon, and that P1 and P2 may be in a 3:2 resonance with each other. Further investigation of these possible resonances, or slight deviations from these resonances, will undoubtedly provide significant insights into the dynamical evolution of the Pluto system.

After finding the orbital solutions described above, we re-examined

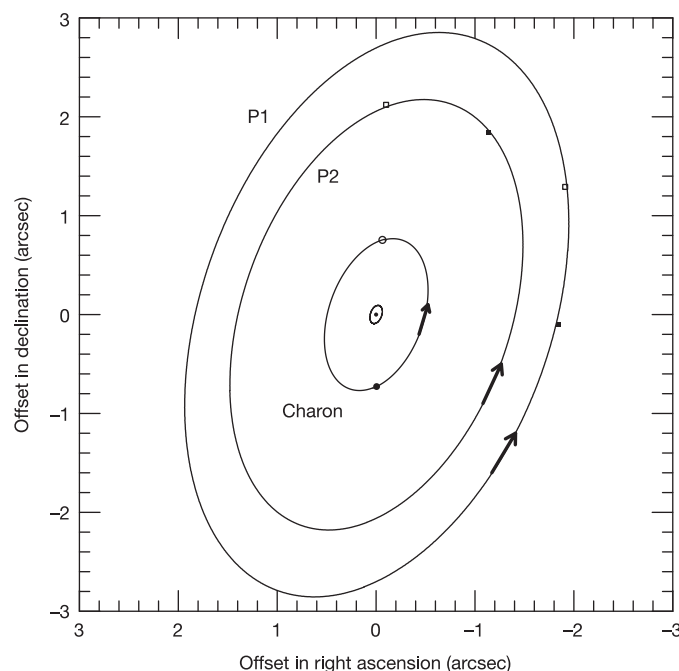


Figure 2 | Preliminary orbits for P1 and P2, assuming that they are circular and in the same plane as Charon's orbit. The barycentre of the system is the dot in the centre, Pluto's orbit is the smallest ellipse, Charon's orbit is the next ellipse (its positions on 15 May and 18 May are indicated by the filled and open circles, respectively), an orbit that is consistent with P2's measured positions is next, followed by an orbit that is consistent with P1's measured positions. For the last two cases, the filled squares are the observed positions on 15 May, and the open squares are the observed positions on 18 May. The uncertainties (1σ) in the measured positions are approximately the size of the symbols. At Pluto's distance, 1 arcsec projects to 21,800 km. Charon's orbital plane is inclined by ~36° relative to the line-of-sight, which means that the circular orbits of the satellites appear elliptical when projected on the plane of the sky. All four objects orbit the barycentre in an anticlockwise direction in this view, and all positions are displayed in the J2000 reference frame with celestial north up and east to the left.

some Hubble Space Telescope data of Pluto taken in 2002 under another programme (M.W.B., unpublished work). That programme was optimized to study the surfaces of Pluto and Charon and used much shorter exposure times (6 s for a V-band filter and 12 s for a B-band-filter) than employed in our programme. Nevertheless, by combining all the images obtained on a given date in the same filter, a limiting sensitivity could be achieved that is very close to the observed magnitudes of P1 and P2 (see Table 1). On the observation date (UT 2002 June 14) with the best geometry (lowest phase angle and smallest geocentric distance), two objects are weakly detected ($S/N \approx 4$) along the predicted orbital paths of P1 and P2 in both the V-band and B-band images, providing independent evidence that P1 and P2 are indeed satellites of Pluto.

Our photometry results are summarized in Table 1. On 15 May P2 has a visual magnitude (V) of 23.38 ± 0.17 , and on 18 May P1 has $V = 22.93 \pm 0.12$. P1 is too close to a diffraction spike for accurate photometry on 15 May, but the brightness then seems consistent with the value observed on 18 May. P2 is too close to a diffraction spike for accurate photometry on 18 May, but the brightness then seems consistent with the value observed on 15 May. Small bodies are often highly irregular in shape, which results in significant temporal variation in brightness as the object rotates. Whether this is true for P1 and P2 must await future investigations.

If the geometric albedo is known, the size of an object can be calculated from its magnitude using a standard relation¹⁰. For a geometric albedo of 0.04 (that is, comet-like¹¹), P1 has a diameter of 167 ± 10 km, and P2 has a diameter of 137 ± 11 km. If the albedo is 0.35 (that is, Charon-like¹²), the diameters are 61 ± 4 km for P1 and 46 ± 4 km for P2. Both satellites are tiny compared to Pluto and Charon, which have diameters of $2,328 \pm 42$ km (refs 2, 3) and $1,207.2 \pm 2.8$ km (ref. 4), respectively. (Using other data from the Charon occultation, Gulbis *et al.*¹⁷ derive a diameter of $1,212 \pm 16$ km.) The masses of P1 and P2 account for less than 5×10^{-4} of the mass of the Pluto system, assuming that the densities of P1 and P2 are similar to, or smaller than, Pluto's density.

We also searched the entire ACS/WFC field-of-view ($202 \text{ arcsec} \times 202 \text{ arcsec}$) for other satellites, but none were found down to a limiting magnitude of $V \leq 26.2$ (90% confidence limit) for the region between 5 and 100 arcsec from Pluto¹³. Thus, the Pluto system of satellites appears to be very compact.

The implications of the discovery of P1 and P2 for the origin and evolution of the Pluto system, and for the satellite formation process in the Kuiper belt, are discussed in a companion paper¹⁴.

Received 18 November; accepted 22 December 2005.

- Christy, J. W. & Harrington, R. S. The satellite of Pluto. *Astron. J.* **83**, 1005–1008 (1978).
- Buie, M. W., Tholen, D. J. & Horne, K. Albedo maps of Pluto and Charon—Initial mutual event results. *Icarus* **97**, 211–227 (1992).
- Young, L. A. *Bulk Properties and Atmospheric Structure of Pluto and Charon*. PhD thesis, MIT (1994).
- Sicardy, B. *et al.* Charon's size and an upper limit on its atmosphere from a stellar occultation. *Nature* **439**, 52–54 (2006).
- Stern, S. A., Parker, J. W., Fesen, R. A., Barker, E. S. & Trafton, L. M. A search for distant satellites of Pluto. *Icarus* **94**, 246–249 (1991).
- Stern, S. A., Parker, J. W., Duncan, M. J., Snowball, J. C. J. & Levison, H. F. Dynamical and observational constraints on satellites in the inner Pluto–Charon system. *Icarus* **108**, 234–242 (1994).
- Stern, S. A. Does the Pluto system contain additional satellites? *Lunar Planet. Inst. Conf. XXXIV*, abstr. 1106 (2003).
- Hartig, G. F., Krist, J. E., Martel, A. R., Ford, H. C., Illingworth, G. D.. *Proc. SPIE* **4854**, 532–543 (2003).
- Bernstein, G. M. *et al.* The size distribution of trans-Neptunian bodies. *Astron. J.* **128**, 1364–1390 (2004).
- Russell, H. N. On the albedo of the planets and their satellites. *Astrophys. J.* **43**, 173–195 (1916).
- Lamy, P. L., Toth, I., Fernandez, Y. R. & Weaver, H. A. in *Comets II* (eds Festou, M. C., Keller, H. U. & Weaver, H. A.) 223–264 (Univ. Arizona Press, Tucson, 2004).
- Marcialis, R. L. *et al.* The albedos of Pluto and Charon—Wavelength dependence. *Astron. J.* **103**, 1389–1394 (1992).
- Steffl, A. J. *et al.* New constraints on additional satellites of the Pluto system. *Astron. J.* (submitted); preprint at (<http://arXiv.org/astro-ph/0511837>).
- Stern, S. A. *et al.* A giant impact origin for Pluto's small moons and satellite multiplicity in the Kuiper belt. *Nature* doi:10.1038/nature04548 (this issue).
- Sirianni, M. *et al.* The photometric performance and calibration of the Hubble Space Telescope Advanced Camera for Surveys. *Publ. Astron. Soc. Pacif.* **117**, 1049–1112 (2005).
- Allen, C. W. *Astrophysical Quantities* 3rd edn (Atholone, London, 1976).
- Gulbis, A. A. S. *et al.* Charon's radius and atmospheric constraints from observations of a stellar occultation. *Nature* **439**, 48–51 (2006).

Acknowledgements We thank G. Hartig for discussions of the ACS optical performance and for examining the images discussed here. We thank the Directors and staff at the Keck, Very Large Telescope, and Gemini observatories for their efforts in attempting ground-based recoveries of these new satellites under non-optimal conditions. We thank the Director and staff of the Space Telescope Science Institute (STScI) for their support of the Hubble Space Telescope observations. Support for this work was provided by NASA through a grant from the STScI, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.A.W. (hal.weaver@jhuapl.edu).

LETTERS

A giant impact origin for Pluto's small moons and satellite multiplicity in the Kuiper belt

S. A. Stern¹, H. A. Weaver², A. J. Steffl¹, M. J. Mutchler³, W. J. Merline¹, M. W. Buie⁴, E. F. Young¹, L. A. Young¹ & J. R. Spencer¹

The two newly discovered¹ satellites of Pluto (P1 and P2) have masses that are small compared to both Pluto and Charon—that is, between 5×10^{-4} and 1×10^{-5} of Pluto's mass, and between 5×10^{-3} and 1×10^{-4} of Charon's mass. This discovery, combined with the constraints on the absence of more distant satellites of Pluto², reveal that Pluto and its moons comprise an unusual, highly compact, quadruple system. These facts naturally raise the question of how this puzzling satellite system came to be. Here we show that P1 and P2's proximity to Pluto and Charon, the fact that P1 and P2 are on near-circular orbits in the same plane as Pluto's large satellite Charon¹, along with their apparent locations in or near high-order mean-motion resonances, all probably result from their being constructed from collisional ejecta that originated from the Pluto–Charon formation event. We also argue that dust–ice rings of variable optical depths form sporadically in the Pluto system, and that rich satellite systems may be found—perhaps frequently—around other large Kuiper belt objects.

The orbits of P1 and P2 reveal that Pluto's satellite system is both largely empty and highly compact (Fig. 1). All three of Pluto's known satellites orbit in the inner $\sim 3\%$ of Pluto's satellite prograde orbit stability radius³, which extends outward to 2.2×10^6 km from Pluto. Outside the three satellite orbits, the system appears to be devoid of other bodies².

We calculated the characteristic tidal e-folding spin-down time⁴, $T_{\text{spin-down}}^{\text{tidal}}$, for these bodies, to evaluate whether they should be expected to have rotational periods similar to their weeks-long orbital periods. We assumed standard values for the mass and radius of Pluto⁵, and minimal masses of P1 and P2. For P1, assuming a rigidity like H₂O–ice and a dissipation factor $Q = 100$, we found $T_{\text{spin-down}}^{\text{tidal}} \approx 10^{11}$ yr in its current orbit. Making the same assumptions for P2, we found $T_{\text{spin-down}}^{\text{tidal}} \approx 10^{12}$ yr in its current orbit. It is therefore clear that the characteristic spin-down times for both P1 and P2 are expected to significantly exceed the 4.5 Gyr age of the Solar System. As a result, P1 and P2 are not expected to be in synchronous rotation with Pluto unless they previously orbited much closer to Pluto, where the spin-down time is decreased by orders of magnitude. If P1 or P2 are discovered to be spin–orbit synchronized, it would therefore suggest that these satellites formerly spent some considerable time closer to Pluto and then subsequently migrated outward as Charon and Pluto exchanged orbital and spin angular momentum to reach their current tidal equilibrium state. We return to this point later when discussing the origin of the system.

First, however, we develop some collisional considerations. Studies of the collisional environment of the present-day Kuiper belt acting on Pluto–Charon and Kuiper belt bodies of smaller sizes revealed^{6,7}

that the critical size boundary for catastrophic breakup over the past 4 Gyr occurs at diameters of ~ 4 km. P1 and P2 are large compared to this critical size scale for catastrophic disruption in the Kuiper belt. P1 and P2 are thus likely to be ancient bodies originally formed during the same era as Pluto and Charon, and are unlikely to have been subsequently disrupted and re-accreted in the past few Gyr.

Collisional studies have also revealed⁶ that in the present day Kuiper belt, the cumulative fraction of the surface cratered by all 8 m diameter and larger Kuiper belt object (KBO) impactors ranges from $\sim 7\%$ to $\sim 32\%$ for bodies on orbits approximately like Pluto's. This does not include the additional surface area covered by ejecta blankets, which would increase this by a factor of 2 to 4, nor does it take into account higher cratering rates in the ancient Kuiper belt, before its mass depletion. Even for these conservative assumptions, we can predict that the surfaces of P1 and P2 will be significantly cratered.

Characteristic collisional velocities for Kuiper belt impactors onto these satellites are in the range $1\text{--}2 \text{ km s}^{-1}$. It has been demonstrated⁶ that objects of the size class of P1 and P2 in the Kuiper belt have probably lost some 10–20% of their mass to impact erosion. We conclude from this that the present day sizes and mass of P1 and P2 are unlikely to be very different from their sizes and masses at the time of their formation.

The characteristic ejecta velocity resulting from collisions onto these satellites should be of order 1–10% of the speed of Kuiper belt debris impactors, or $10\text{--}100 \text{ m s}^{-1}$. At these speeds, collisional ejecta fragments will escape the satellites themselves but generally remain trapped in orbit about Pluto. This is in contrast to the situation obtaining at Charon (with its $\sim 500 \text{ m s}^{-1}$ escape velocity) because most collisional ejecta falls back onto Charon's surface, and does not reach orbit about Pluto. As such, the bombardment of P1 and P2 by small Kuiper belt debris almost certainly generates faint, dusty ice particle rings around Pluto, with time-variable optical depth.

A crude estimate of the approximate optical depth of these rings can be derived by assuming that 10% of the mass of these satellites may have been eroded from them over time⁶. If we then adopt the conservative assumptions of (1) minimum estimated satellite masses¹ for P1 and P2, (2) a mean lifetime for ejected particles of 10^5 yr (that is, an order of magnitude shorter than the estimated lifetime against erosion and sublimation of Kuiper belt dust particles⁸), (3) that only 10^{-4} of the debris is in micrometre-sized particles, (4) that the ring particles have 1 g cm^{-3} density, and (5) a (conservative) characteristic width spanning the entire separation between P1 and P2, we derive a characteristic ring optical depth estimate of $\tau = 5 \times 10^{-6}$. This is comparable to the optical depth of Jupiter's tenuous ring system. This estimate is only very approximate,

¹Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, Colorado 80302, USA. ²The Johns Hopkins Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, Maryland 20723, USA. ³Space Telescope Science Institute, 3700 San Martin Road, Baltimore, Maryland 21218, USA. ⁴Lowell Observatory, 1400 West Mars Hill Road, Flagstaff, Arizona 86001, USA.

but from it we conclude that Pluto can transiently possess dust rings as a result of the stochastic bombardment of P1 and P2 by small Kuiper belt debris.

Now consider how P1 and P2 may have formed. Pluto's satellite Charon is half of Pluto's diameter, and has a specific angular momentum so high that there is broad agreement that the pair was generated via a giant collision with an ancient impactor^{9–12}. But what is the origin of P1 and P2, two remarkably smaller satellites exterior to Charon?

P1 and P2's proximity to Pluto and to Charon, their apparent locations in or near high-order mean-motion resonances, and in the plane of Charon's orbit¹, together present strong challenges to any assumed capture origin, but naturally suggest a formation in association with the giant impact origin of Charon. We therefore suggest that P1 and P2 are, like Charon, likely to be constructed of material ejected into orbit about Pluto as a result of the Charon-forming impact event.

This hypothesis is further supported by the circular or near-circular orbits of P1 and P2. We elaborate on this case by estimating the characteristic e-folding eccentricity decay time⁴:

$$\tau = e / (de/dt) = 2Q_s / [21\mu n(R_s/a_s)^5 k_2]$$

Here e is eccentricity, Q_s is the dissipation coefficient for the satellite, μ is the satellite mass ratio relative to its primary (Pluto in this case), n is the orbital mean motion of the satellite, R_s is the satellite's radius,

a_s is the satellite's orbital semi-major axis, and k_2 is the second degree potential Love number of the satellite.

Adopting the P1 and P2 orbits we reported elsewhere¹, $Q_s = 100$ (considered typical of icy satellites), $k_2 = 0.055$ (appropriate for rigid ice⁴), densities of 2 g cm^{-3} (that is, similar to Pluto and Charon), and assuming that P1 and P2 have their maximum permissible radii¹, we find tidal circularization timescales of 65 Gyr and 500 Gyr, respectively. Thus, it is seen that near their current orbits, or farther out, the eccentricity decay times for P1 and P2 are far too long to damp from high eccentricity capture values to circular orbits in the age of the Solar System unless either the satellite Q_s values are <1 (strengthless rubble piles) and/or unless gas drag assisted any putative eccentricity decay. In contrast, Charon's orbital eccentricity decay timescale is short, $\sim 3 \times 10^6$ yr, and the tidal decay timescale for eccentricity near Pluto's Roche lobe is only of order 10^5 yr. This suggests to us that the circular orbits of P1 and P2 imply that (1) they most probably formed much closer to Pluto, rather than farther out or by capture from heliocentric orbit, and (2) they subsequently evolved outward to their present-day positions during the tidal evolution of Charon to its current orbit.

This said, we note that the very small masses of P1 and P2 relative to Charon beg the question of why so little material would have escaped accumulation into orbiting bodies other than Charon, and therefore, why there are not more small satellites of Pluto. Perhaps other satellites did form, but eventually became dynamically

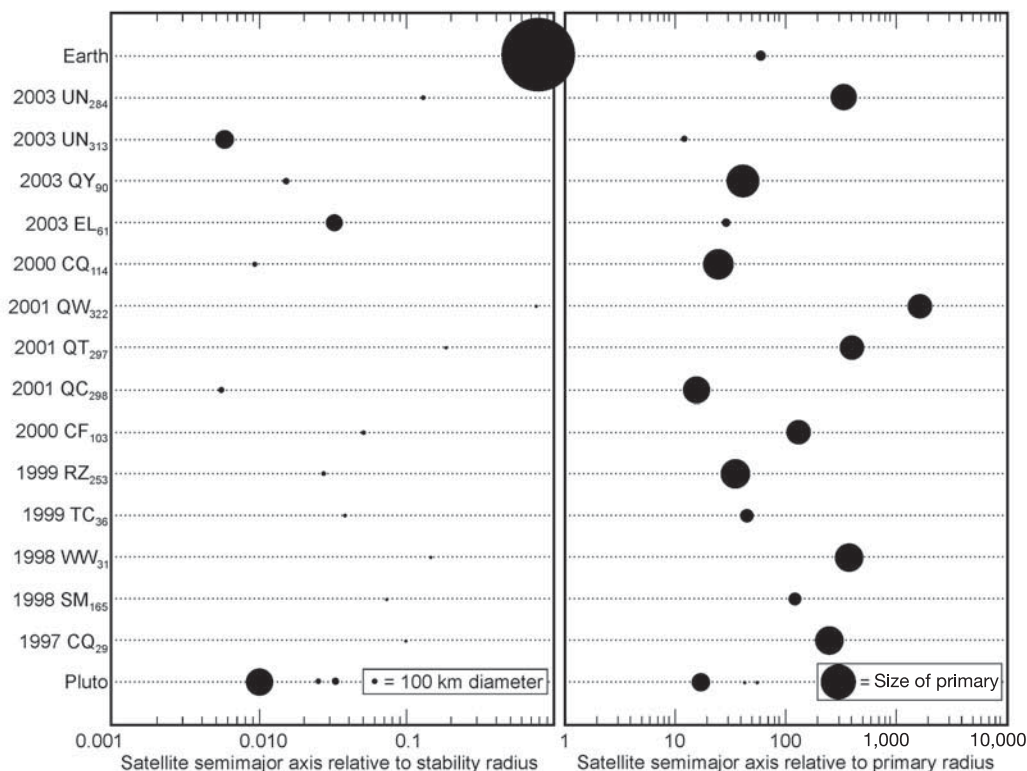


Figure 1 | The architecture of the Pluto system compared to other KBOs with known satellites and to the Earth–Moon system. The orbital distances and sizes of all three satellites in the Pluto system are shown here in comparison to other relatively well characterized KBO–satellite systems, and the Earth–Moon pair. P1 and P2 orbit relatively close to Pluto at distances of $64,700 \pm 850$ km and $49,400 \pm 600$ km, respectively¹. Photometry of these two bodies¹ indicates that their visual magnitudes were $V = 22.93 \pm 0.12$ and 23.38 ± 0.17 , respectively, in mid-May 2005. For an assumed (that is, comet-like) lower limit albedo of 0.04 (as shown), one derives upper limit diameters of 167 ± 10 km for P1 and 137 ± 11 km for P2. If their albedos are as high as 0.35 (that is, like Charon⁵, a reasonable upper limit), then their diameters are only $\sim 61 \pm 4$ km and $\sim 46 \pm 4$ km,

respectively. Pluto apparently has no undiscovered satellites farther out in the system down to objects 40 times fainter than P1 or P2. For this figure, all satellites are assumed to lie at their discovery distance if a formal semi-major axis has not yet been established. The left-hand panel shows satellite sizes on an absolute scale, with orbital distances normalized to the orbital stability zone within which the primary body can retain satellites over long timescales. Masses were computed from sizes assuming a density of $\rho = 2 \text{ g cm}^{-3}$, like Pluto and Charon⁵. The right-hand panel shows the systems with satellite sizes normalized to the radius of the primary in each system (for example, Charon appears larger than the Moon), and their orbital distances in units of the primary's radius; object sizes were computed assuming 4% albedos.

destabilized, resulting in accumulation onto Charon or Pluto; or perhaps there are other, still fainter, satellites that escaped detection below the new Hubble Space Telescope observation² threshold near a visual magnitude of $V = 26.2$.

Finally, it has been estimated that 20%, or more, of the known KBOs have satellites¹³. This suggests that there must be tens of thousands of KBOs with satellites. Given this, and the presence of P1 and P2 orbiting Pluto, we consider it likely that many KBO satellite systems will be revealed to be multiples when examined closely. Figure 1 further illustrates this point, showing the ample space available for minor satellites in the known KBO systems.

We suggest that a natural place to expect multiple satellite systems (and associated rings) would be around those KBOs that possess tightly bound, large satellites reminiscent of binary formation events like the Pluto–Charon pair. Such objects include 1997 CQ₂₉, 1998 SM₁₆₅, 1999 TC₃₆, 2003 UB₃₁₃ and 2003 EL₆₁. It will also be useful to search for more distant, irregular satellites orbiting KBOs to determine whether less-compact (for example, capture-related) architectures also exist among KBOs with satellite systems.

Received 17 November; accepted 20 December 2005.

1. Weaver, H. A. *et al.* Discovery of two new satellites of Pluto. *Nature* doi:10.1038/nature04547 (this issue).
2. Steffl, A. J. *et al.* Brightness constraints on additional satellites in the Pluto system. *Astron. J.* (submitted).
3. Stern, S. A., Parker, J. Wm., Fesen, R. A., Barker, E. S. & Trafton, L. M. A search for distant satellites of Pluto. *Icarus* **94**, 246–249 (1991).
4. Dobrovolskis, A. R., Peale, S. J. & Harris, A. W. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 159–190 (Univ. Arizona Press, Tucson, 1997).
5. Tholen, D. J. & Buie, M. W. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 193–219 (Univ. Arizona Press, Tucson, 1997).
6. Durda, D. D. & Stern, S. A. Collision rates in the present-day Kuiper Belt and Centaur regions: Applications to surface activation and modification on comets, Kuiper Belt objects, Centaurs, and Pluto–Charon. *Icarus* **145**, 220–229 (2000).
7. Stern, S. A. Collisional time scales in the Kuiper disk and their implication. *Astron. J.* **110**, 856–860 (1995).
8. Mukai, T. *et al.* Small bodies and dust in the outer solar system. *Adv. Space Res.* **34**, 172–178 (2004).
9. McKinnon, W. B. On the origin of Triton and Pluto. *Nature* **311**, 355–358 (1984).
10. McKinnon, W. B. On the origin of the Pluto–Charon binary. *Astrophys. J.* **344**, L41–L44 (1989).
11. Stern, S. A., McKinnon, W. B. & Lunine, J. I. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 605–663 (Univ. Arizona Press, Tucson, 1997).
12. Canup, R. M. Giant impact origin of Pluto–Charon. *Science* **307**, 546–550 (2005).
13. Stephens, D. C. & Noll, K. S. The high fraction of binaries in the cold classical Kuiper Belt. *Astron. J.* **131**, 1142–1148 (2006).

Acknowledgements We thank W. McKinnon, R. Canup and W. Ward for reading and commenting on this manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.A.S. (astern@swri.edu).

Counterfactual quantum computation through quantum interrogation

Onur Hosten¹, Matthew T. Rakher^{1†}, Julio T. Barreiro¹, Nicholas A. Peters¹ & Paul G. Kwiat¹

The logic underlying the coherent nature of quantum information processing often deviates from intuitive reasoning, leading to surprising effects. Counterfactual computation constitutes a striking example: the potential outcome of a quantum computation can be inferred, even if the computer is not run¹. Relying on similar arguments to interaction-free measurements² (or quantum interrogation³), counterfactual computation is accomplished by putting the computer in a superposition of ‘running’ and ‘not running’ states, and then interfering the two histories. Conditional on the as-yet-unknown outcome of the computation, it is sometimes possible to counterfactually infer information about the solution. Here we demonstrate counterfactual computation, implementing Grover’s search algorithm with an all-optical approach⁴. It was believed that the overall probability of such counterfactual inference is intrinsically limited^{1,5}, so that it could not perform better on average than random guesses. However, using a novel ‘chained’ version of the quantum Zeno effect⁶, we show how to boost the counterfactual inference probability to unity, thereby beating the random guessing limit. Our methods are general and apply to any physical system, as illustrated by a discussion of trapped-ion systems. Finally, we briefly show that, in certain circumstances, counterfactual computation can eliminate errors induced by decoherence.

The essential feature of Grover’s algorithm is that an amplitude-enhancement technique transfers the amplitude from a uniform database distribution to a particular element ‘marked’ by an ‘Oracle-type’ processor^{4,7}. For instance, consider a database of four elements with input state $|00\rangle$. At the end of the algorithm, if the marked element (ME) is no. 1, no. 2, no. 3 or no. 4, the final state of the readout qubits is $|00\rangle$, $|01\rangle$, $|10\rangle$ or $|11\rangle$, respectively.

Counterfactual computation (CFC) conceptually proceeds as follows: (0) the initial state of the computer can be written as $|\psi_{\text{in}}\rangle = |\text{Off}\rangle|00\rangle$, the $|\text{Off}/\text{On}\rangle$ qubit being the ‘operating switch’; (1) apply a $\frac{\pi}{2}$ -rotation ($R(\frac{\pi}{2}) : |\text{Off}\rangle \rightarrow \frac{|\text{Off}\rangle + |\text{On}\rangle}{\sqrt{2}}$ and $|\text{On}\rangle \rightarrow \frac{-|\text{Off}\rangle + |\text{On}\rangle}{\sqrt{2}}$) to the ‘switch’; (2) run the algorithm if the ‘switch’ is ‘On’; (3) apply $R(\frac{\pi}{2})$ to the ‘switch’ only if the registers are in state $|00\rangle$. If the ME is no. 1, the effect is:

$$\begin{aligned} |\text{Off}\rangle|00\rangle &\xrightarrow{R} \frac{|\text{Off}\rangle + |\text{On}\rangle}{\sqrt{2}}|00\rangle \xrightarrow{\text{Grover}} \\ &\frac{|\text{Off}\rangle|00\rangle + |\text{On}\rangle|00\rangle}{\sqrt{2}} \xrightarrow{R} |\text{On}\rangle|00\rangle \end{aligned} \quad (1)$$

All the amplitude ends up in state $|\text{On}\rangle|00\rangle$, with equal amplitudes constructively interfering from histories with the computer ‘running’ and ‘not running’.

For other MEs (with $xy = 01, 10$ or 11), we have:

$$\begin{aligned} |\text{Off}\rangle|00\rangle &\xrightarrow{R} \frac{|\text{Off}\rangle + |\text{On}\rangle}{\sqrt{2}}|00\rangle \xrightarrow{\text{Grover}} \frac{|\text{Off}\rangle|00\rangle + |\text{On}\rangle|xy\rangle}{\sqrt{2}} \xrightarrow{R} \\ &\frac{|\text{Off}\rangle + |\text{On}\rangle}{2}|00\rangle + \frac{1}{\sqrt{2}}|\text{On}\rangle|xy\rangle \end{aligned} \quad (2)$$

Now, there is a $\frac{1}{4}$ probability to measure the final state $|\text{Off}\rangle|00\rangle$, from which we conclude that the ME is not no. 1 (as this term did not appear when ME = no. 1). There is no amplitude from a history with the computer ‘running’ in this outcome; therefore, we can conclude that the ME is not no. 1 without the computer ‘running’. (Later we describe a method to counterfactually determine the actual outcome.)

An optical realization of CFC is shown in Fig. 1 as an interferometer. We use the optical circuit in ref. 8 for Grover’s algorithm (shown as a black box in Fig. 1), with slight changes to improve performance (see Supplementary Information). This optical circuit takes in a single photon in path a with H (horizontal) polarization, that is, $|aH\rangle$. The output path a or b and polarization H or V (vertical) of the photon depend on the ME: no. 1 $\rightarrow |aH\rangle$, no. 2 $\rightarrow |aV\rangle$, no. 3 $\rightarrow |bH\rangle$ and no. 4 $\rightarrow |bV\rangle$. Such single-photon encoding, though not scalable, suffices for our pedagogical purpose. Figure 2a shows that the algorithm as realized operates quite precisely, with an average error of less than 2.6%.

To realize CFC, the path lengths of the interferometer in Fig. 1 are adjusted to give destructive interference at detector D_1 when the ME is no. 1, that is, if the photon traverses both paths indistinguishably, it is definitely detected at D_5 (Fig. 1). If the ME is not no. 1, there is no amplitude exiting the search algorithm in the $|aH\rangle$ mode, and no amplitude from the algorithm can reach the second beam-splitter (BS), thus eliminating the destructive interference at D_1 . Therefore, this time there is a $\frac{1}{4}$ probability of detecting the photon at D_1 , indicating that the ME is not no. 1, even though the photon came from the algorithm-free path (that is, the computer did not run). Other possibilities are that with probability $\frac{1}{2}$ we find out the answer but the photon comes from the algorithm (D_2 , D_3 or D_4), and with probability $\frac{1}{4}$ we detect the photon at D_5 (after it has travelled the algorithm-free path) leading to no definite information (as a detection also occurs at D_5 if ME = no. 1). The efficiency of CFC can be quantified as $\eta = P_{D_1}/(P_{D_1} + P_{\text{Grover}}) = 1/3$, where P_{D_1} is the probability of measuring the photon at D_1 , a CFC; and P_{Grover} is the probability of the photon passing through the algorithm.

Experimentally, we used an equivalent but more stable interferometer configuration (see Supplementary Information); our CFC system performed as indicated in Fig. 2b, with an efficiency $\langle\eta\rangle = 0.319 \pm 0.009$ (averaged over ME $\neq$ no. 1). The interrogation interferometer had an intrinsic visibility of 97%, implying imperfect

¹Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [†]Present address: Department of Physics, University of California Santa Barbara, Santa Barbara, California 93106, USA.

destructive interference at the D_1 output when the ME was no. 1. A detection at the D_1 output can thus no longer indicate a CFC with 100% certainty. We characterize this feature in terms of P_{Succ} , the probability of a successful CFC upon a detection at the interrogation detector D_1 : $P_{\text{Succ}|no.i} = P_{D_1|no.i} / (P_{D_1|no.1} + P_{D_1|no.i})$. Here the probabilities are conditional on the ME ($i = 2, 3, 4$). In the experiment we achieved $\langle P_{\text{Succ}} \rangle = 0.943 \pm 0.009$.

If one replaces the first (second) 50/50 beam splitter by a highly reflecting (transmitting) one, reducing the amplitude passing through the algorithm, η can be increased⁹ to a maximum value of $\frac{1}{2}$. We tested this (here using attenuated coherent states) with a 5/95 BS, achieving $\langle \eta \rangle = 0.472 \pm 0.007$ (0.487 theoretically), with a slight decrease in the CFC success: $\langle P_{\text{Succ}} \rangle = 0.877 \pm 0.009$. (See Supplementary Information for a modified version of CFC that interrogates all database elements simultaneously to determine the ME itself; however, this modified version is also limited to $\eta = \frac{1}{2}$).

The efficiency can be increased from $\frac{1}{2}$ to 1 using the quantum Zeno effect¹, just as it was used for quantum interrogation^{3,9}. (Before proceeding, we note that our approach should not be confused with the interesting proposal to combine two-photon absorption and the quantum Zeno effect to enable efficient optical quantum computation¹⁰, nor with the suggestion to use quantum interrogation inside Grover's search algorithm¹¹.) Specifically, the switch is rotated successively in small steps ($R(2\theta)$: $|Off\rangle \rightarrow \cos\theta|Off\rangle + \sin\theta|On\rangle$ and $|On\rangle \rightarrow -\sin\theta|Off\rangle + \cos\theta|On\rangle$; $\theta = \frac{\pi}{2N}$, integer $N \gg 1$) from state $|Off\rangle$ to $|On\rangle$, and the output registers are monitored at each step. If the ME is no. 1, measurements on the output registers result in $|00\rangle$ without affecting the evolution due to rotations, leaving the system in $|On\rangle|00\rangle$ after N rotations. However, if the ME is not no. 1, then the system evolves as:

$$|Off\rangle|00\rangle \xrightarrow{R} (\cos\theta|Off\rangle + \sin\theta|On\rangle)|00\rangle \xrightarrow{\text{Grover}} \cos\theta|Off\rangle|00\rangle + \sin\theta|On\rangle|xy\rangle \xrightarrow{\text{Measure}} \approx \cos\theta|Off\rangle|00\rangle \quad (3)$$

The measurement on the output registers results in state $|00\rangle$, leaving the computer in the $|Off\rangle$ state with probability $\cos^2(\frac{\pi}{2N})$. After a total of N cycles, the probability of finding the system in state $|Off\rangle|00\rangle$ is $\cos^{2N}(\frac{\pi}{2N})$, which tends to 1 as $N \rightarrow \infty$. Thus, if the ME is not no. 1, the final state is $|Off\rangle|00\rangle$ without the computer running; and if the ME is no. 1, the final state is $|On\rangle|00\rangle$, that is, this time the computer runs. Note that at best one can exclude a single value of the ME with these techniques.

Now we describe a novel 'chained Zeno effect' that will permit us to counterfactually determine the actual ME with $\eta \rightarrow 1$. The strategy is to place the above 'Zeno scheme' inside another one, to avoid the computer running even if the ME is no. 1. For this purpose we use a third 'switch' state $|Off'\rangle$, in addition to $|Off\rangle$ and $|On\rangle$, and

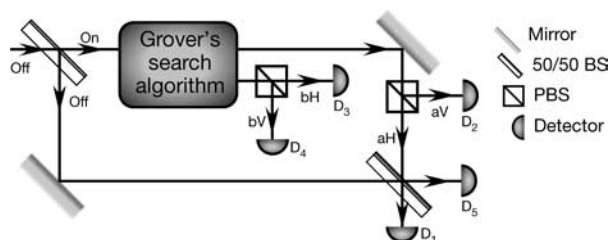


Figure 1 | An optical realization of counterfactual computation. By means of a 50/50 beam splitter (BS) (which serves as a $\frac{\pi}{2}$ -rotation), an H-polarized single photon is put in a superposition of passing and not passing through the algorithm, encoding the 'operating switch' in different spatial modes, 'On' and 'Off'. Then on a second 50/50 BS, the two histories are interfered only if the photon after the algorithm is in the mode $|aH\rangle$. The modes $|aH\rangle$ and $|aV\rangle$ are distinguished via a polarizing beam splitter (PBS) which transmits H and reflects V.

define an additional rotation R' to couple $|Off'\rangle$ to $|Off\rangle$ ($R'(2\theta')$: $|Off'\rangle \rightarrow \cos\theta'|Off'\rangle + \sin\theta'|Off\rangle$ and $|Off\rangle \rightarrow -\sin\theta'|Off'\rangle + \cos\theta'|Off\rangle$; $\theta' = \frac{\pi}{2N'}$). A possible optical implementation of the technique is shown in Fig. 3a. A single photon starts in cavity 'Off'. Using active optical elements (Pockels cells), a small amount of amplitude is exchanged between 'Off' and 'Off' via BS₁ (see Methods for details). The small amplitude then performs the high-efficiency quantum Zeno cycles described above (N times between cavities 'Off' and 'On'). If the ME $\neq$ no. 1, the small amplitude component effectively stays in cavity 'Off'. But if ME = no. 1, then first, all the small amplitude component transfers to cavity 'On', and then, via the Pockels cell in cavity 'On', is actively absorbed at A_1 . Now the entire procedure starting with the amplitude exchange between 'Off' and 'Off' is repeated, a total of N' times. At the end of $N' \times N$ total cycles, if ME = no. 1 ($\neq$ no. 1) then the photon will be measured in cavity 'Off' ('Off') with probability approaching one (Fig. 3e) as $N' \rightarrow \infty$ ($\frac{N'}{N} \rightarrow \infty$). In neither of these cases does the computer 'run'. One can then re-interrogate for the other elements one by one—thereby identifying the ME counterfactually—by changing the connections to the algorithm (Fig. 3b). Curves in Fig. 3e characterize a lossless system; for large N' , N , even small losses become detrimental³, limiting the achievable performance in any real system.

The necessity to interrogate database elements one by one would negate the quantum speed-up advantage. However, it is possible to circumvent this by interrogating the logical value of each qubit one by one instead. To do this, we need to implement both the search algorithm and its adjoint—which undoes the search—and we need to perform the measurements (giving rise to the Zeno effect) between the algorithm and its adjoint. The first (second) optical circuit in Fig. 3c can be used in cavity 'On', to interrogate for the logical value of the first (second) qubit. If the value of the first (second) qubit is 'a' (H), then, at the end of the cycles, the photon will be measured in cavity 'Off'; likewise if the value is 'b' (V), then the photon will be

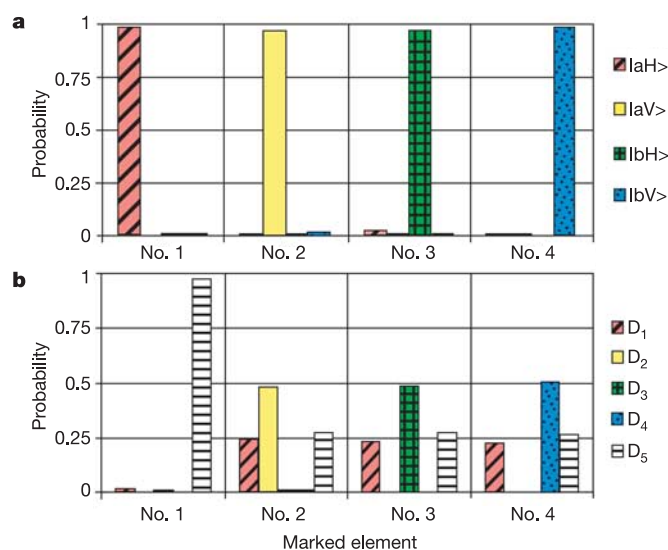


Figure 2 | Experimentally determined probabilities for the output state of 670-nm single photons conditionally prepared through downconversion¹⁷. **a**, The performance of the search algorithm (see Supplementary Information for details). The total probability of finding the photon in any incorrect output port is 2.6% (averaged over MEs); thus, the ME could be accurately determined with a single photon passing through the algorithm. **b**, Output state probabilities for the set-up in Fig. 1. Grid lines closest to data points represent theory. We attribute slight deviations from theory to imperfect beam splitting ratios, imperfect mode matching (apparently from wavefront distortions in various elements), and imperfect path-length balancing.

measured in cavity 'Off'. In neither case does the computer (or its adjoint) run. (See Supplementary Information for more details, including the quantum circuit diagram of this method.)

The physical implementation of our proposals is not limited to optical systems. Consider qubits encoded using hyperfine ground states ($|0\rangle$ and $|1\rangle$) of two trapped ions. In the simplest scheme (the analogue of Fig. 1), the ions would start in trap 'Off' ($|Off\rangle|00\rangle$). The atoms can be prepared in an equal superposition of both atoms being in trap 'Off' or being in another trap 'On' (see Methods). Then, a measurement on the atoms can be performed, causing at least one of them to fluoresce if the ME $\neq$ no. 1 but leaving the system undisturbed if ME = no. 1. After interfering the two histories, if ME = no. 1 the atoms end up in trap 'On'. If instead ME $\neq$ no. 1, then there is a $\frac{1}{4}$ probability of finding the atoms in trap 'Off' without observing any fluorescence during the measurement process above (which means that the algorithm did not run), from which we could counterfactually conclude that the ME $\neq$ no. 1. In the 'chained Zeno' version, one would repeatedly look to see if the algorithm had run; but the atoms never appear in trap 'On', and the final location of atoms (trap 'Off' or 'Off') would counterfactually reveal the results.

To the best of our knowledge, we have achieved the most accurate (2.6% error) realization to date of Grover's search algorithm, albeit with a non-scalable single-photon implementation. Using this set-up, we made the first proof-of-principle demonstration of CFC, inferring that a particular element was not the answer to the computation, even though the computer did not run (with efficiency $\eta \approx 0.319$). Then, we showed in principle how to obtain complete information from the algorithm unconditionally, without the algorithm ever running, using a 'chained' Zeno effect. In fact, these CFC methods still work if the algorithm were to output non-orthogonal states as answers (for example, Grover's algorithm in a larger

database), or if the 'Oracle' of the search algorithm were to be in a quantum superposition of marking different elements (see Supplementary Information).

Decoherence in quantum computing occurs owing to coupling to uncontrolled degrees of freedom—the environment—and results in incomplete interference between qubit states, leading to errors. In CFC, although the algorithm does not run, it still needs to be ready to run correctly, for all the critical interference effects to occur (for example, for the case of trapped ion computation, all the laser pulses for state manipulation still need to be sent into the trap, even if after the fact there was no ion there). For this reason, one might conclude that CFC would still necessarily be subject to decoherence-induced errors. In the Methods section, however, we show that by using a variant of the 'chained' Zeno CFC, in certain circumstances it is possible to protect computations against decoherence. When this scheme succeeds (that is, when the photon does not get absorbed; Fig. 3a and d), the total probability amplitude that has actually run through the algorithm is very small ($\ll 1$, though not zero). Consequently, coupling to the environment is small and decoherence-induced errors are suppressed. It remains to be seen whether the success probability can be made arbitrarily close to unity regardless of the amount of error. Nevertheless, by slightly modifying our approach, it may be possible to design error-suppressing counterfactual qubit gates, for use in scalable quantum computing architectures.

METHODS

CFC using the 'chained Zeno effect'. Using a polarizing BS and Pockels cells, a single photon is switched into cavity 'Off' (Fig. 3a). By firing the Pockels cells at appropriate times, cavities 'Off' and 'Off' are allowed a single coherent amplitude exchange via BS₁, after which the cavities are isolated again. While the vast majority of the amplitude is cycling in cavity 'Off', the small amount of amplitude that has been allowed to leak into cavity 'Off' performs the high-efficiency quantum Zeno cycles described in the text (N times between cavities 'Off' and 'On'). At each step, the amplitude exiting the algorithm is directed to an absorber (A₂ or A₃) if the ME is different from no. 1, effectively projecting the leaked amplitude into cavity 'Off', assuming $N \gg 1$. In contrast, if the ME is no. 1, at the end of N cycles all of the small amplitude component has coherently moved into cavity 'On'; when the Pockels cell in this cavity is fired to terminate the amplitude at A₁, no amplitude is left in either cavity 'Off' or 'On'. In fact, this effectively projects the state into cavity 'Off', assuming $N' \gg 1$.

Now we once again temporarily couple cavities 'Off' and 'Off', and repeat the 'Off'–'On' (inner) cycling another N times. The procedure ends after repeating this entire procedure (outer cycle) N' times, for a total of $N' \times N$ inner cycles. Upon successful operation, if the ME is no. 1 (not no. 1), the photon ends up in cavity 'Off' ('Off').

Error suppression. In our theoretical model, decoherence causes depolarization. Assume that, with probability ϵ , the search algorithm becomes entangled with the environment, and outputs $(1 - \frac{3\epsilon}{4})|00\rangle\langle 00| + \frac{\epsilon}{4}|01\rangle\langle 01| + \frac{\epsilon}{4}|10\rangle\langle 10| + \frac{\epsilon}{4}|11\rangle\langle 11|$ instead of $|00\rangle\langle 00|$, outputs $\frac{\epsilon}{4}|00\rangle\langle 00| + (1 - \frac{3\epsilon}{4})|01\rangle\langle 01| + \frac{\epsilon}{4}|10\rangle\langle 10| + \frac{\epsilon}{4}|11\rangle\langle 11|$ instead of $|01\rangle\langle 01|$, and so on. In an attempt to suppress these errors, we use a variant of the 'qubit-by-qubit' 'chained' Zeno CFC (note that the algorithm adjoint will also introduce errors, making the net probability of entanglement with the environment $2\epsilon - \epsilon^2$). Instead of measurements between the algorithm and its adjoint (Fig. 3c), we apply an extra π -phase shift to the state if the 'operating switch' is $|0\rangle$ and the state of the qubit being interrogated is $|1\rangle$ (Fig. 3d). After the algorithm adjoint we perform a measurement on the output registers to ensure that the state is $|00\rangle$ (that is, absorb the photon if it is not in mode 'aH'), as the search algorithm and its adjoint together should leave the output registers unchanged, that is, $|00\rangle$, in the case of no errors. If the state of the qubit being interrogated is $|1\rangle$, then because of the extra π -phase shift, any amplitude entering the 'On' state is interferometrically directed back to 'Off' (of course degraded by decoherence). But, if the state is $|0\rangle$, then the amplitude coherently builds up in 'On' (also degraded by decoherence), depleting the amplitude in 'Off', in effect inhibiting the coherent flow from 'Off' to 'Off'. At the end, if the qubit being interrogated is $|0\rangle$ ($|1\rangle$), we expect the system to stay in (move to) 'Off' ('Off').

Applying the extra π -phase shifts, instead of measurements, is analogous to the Super-Zeno effect¹² (a bang-bang type control¹³) instead of the Zeno effect, and preserves the quantum state of a system much more efficiently. Optimum calculations ($N' = 70$, $N = 40$; $\frac{N'}{N} \gg 1$ not being necessary for the Super-Zeno

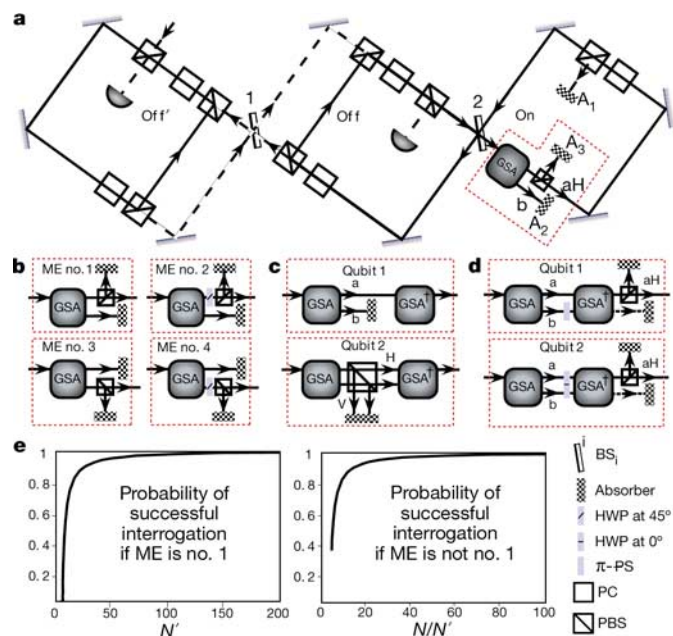


Figure 3 | Proposed set-up for the 'chained Zeno effect'. Three cavities correspond to three states of the switch ($|Off'\rangle$, $|Off\rangle$ and $|On\rangle$), separated by BSs—the rotation operators. Pockels cells (PCs) rotate the polarization by 90° on demand; half-wave plates (HWP) at 45° rotate the polarization by 90° . GSA, Grover's search algorithm. **a**, Interrogation for element no. 1. A₁–A₃ are absorbers. **b**, Settings for the interrogation of different elements. **c**, Set-up configurations for qubit-by-qubit interrogation. GSA[†] undoes the action of GSA. **d**, Configurations for error suppression. π -PS induces a π -phase shift on path b; HWP at 0° induces a π -phase shift on polarization V. **e**, Probability of successful interrogation for the set-up in **a**, as a function of the cycling parameters N' and N (numerically evaluated).

version) show that, for single-pass error probability $\epsilon = 0.05$ ($2\epsilon - \epsilon^2 = 0.0975$), error-suppressing CFC yields the value of the qubit being interrogated 0.944 of the time with an error probability of only $\sim 2 \times 10^{-4}$. The remaining 0.056 ($< 2\epsilon - \epsilon^2$) of the time the system suffers decoherence and fails, and the photon gets absorbed at one of the absorbers in Fig. 3d. Even for a much larger $\epsilon = 0.5$ ($2\epsilon - \epsilon^2 = 0.75$), we still learn the value of the qubit 0.508 of the time, with an error of only ~ 0.035 .

CFC with trapped ions. Consider ions starting in state $|\text{Off}\rangle|00\rangle$. In order to put the computer in a superposition, the atoms can first be prepared in the entangled state $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$ (by applying the rotation $R(\frac{\pi}{2})$ (see main text) to the first qubit, and then applying a controlled-not gate on the qubits); with the aid of appropriate laser beams, a state-dependent spatial force can then be applied to the atoms, resonantly driving and separating the wavepackets of different internal states¹⁴. At this point a new external potential can be introduced, which traps $|0\rangle$ states in trap 'Off', and $|1\rangle$ states in trap 'On': $\frac{|\text{Off}\rangle|00\rangle+|\text{On}\rangle|11\rangle}{\sqrt{2}}$. After flipping the states of the amplitudes in trap 'On', we can apply the algorithm¹⁵ (again using suitable laser pulses) in this trap: $\frac{|\text{Off}\rangle|00\rangle+|\text{On}\rangle|xy\rangle}{\sqrt{2}}$. The states of the atoms can be measured efficiently using a state-selective fluorescence technique¹⁶, which scatters light only from atoms in state $|1\rangle$. But this measurement gives no signal if the output of the algorithm is $|xy=00\rangle$ (that is, if the ME = no. 1), and thus cannot distinguish which trap the ions are in, leaving the state $\frac{|\text{Off}\rangle|00\rangle+|\text{On}\rangle|00\rangle}{\sqrt{2}}$. The amplitudes in the two traps can then be coherently interfered by reversing the amplitude-splitting procedure described above. The extension to high efficiency interrogation can be accomplished by preparing the entangled state $\cos\theta|00\rangle + \sin\theta|11\rangle$ (which only differs in applying the rotation $R(2\theta)$ instead of $R(\frac{\pi}{2})$ to the first qubit).

Received 4 August; accepted 14 December 2005.

- Jozsa, R. in *Lecture Notes in Computer Science* (ed. Williams, C. P.) Vol. 1509, 103–112 (Springer, London, 1998).
- Elitzur, A. C. & Vaidman, L. Quantum-mechanical interaction-free measurements. *Found. Phys.* **23**, 987–997 (1993).
- Kwiat, P. G. *et al.* High-efficiency quantum interrogation measurements via the quantum Zeno effect. *Phys. Rev. Lett.* **83**, 4725–4728 (1999).
- Grover, L. K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79**, 325–328 (1997).
- Mitchison, G. & Jozsa, R. Counterfactual computation. *Proc. R. Soc. Lond. A* **457**, 1175–1193 (2001).
- Misra, B. & Sudarshan, E. C. G. The Zeno's paradox in quantum theory. *J. Math. Phys.* **18**, 756–763 (1977).
- Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* 248–276 (Cambridge Univ. Press, Cambridge, UK, 2000).
- Kwiat, P. G., Mitchell, J. R., Schwindt, P. D. D. & White, A. G. Grover's search algorithm: an optical approach. *J. Mod. Opt.* **47**, 257–266 (2000).
- Kwiat, P. G. *et al.* Interaction-free measurement. *Phys. Rev. Lett.* **74**, 4763–4766 (1995).
- Franson, J. D., Jacobs, B. C. & Pittman, T. B. Quantum computing using single photons and the Zeno effect. *Phys. Rev. A* **70**, 062302 (2004).
- Rudolph, T. & Grover, L. Quantum searching a classical database (or how we learned to stop worrying and love the bomb). Preprint at (<http://arxiv.org/quant-ph/0206066>) (2002).
- Dhar, D., Grover, L. K. & Roy, S. M. Preserving quantum states: A Super-Zeno effect. Preprint at (<http://arxiv.org/quant-ph/0504070>) (2005).
- Viola, L. & Lloyd, S. Dynamical suppression of decoherence in two-state quantum systems. *Phys. Rev. A* **58**, 2733–2744 (1998).
- Monroe, C., Meekhof, D. M., King, B. E. & Wineland, D. J. A "Schrodinger Cat" superposition state of an atom. *Science* **272**, 1131–1136 (1996).
- Brickman, K. A. *et al.* Implementation of Grover's quantum search algorithm in a scalable system. *Phys. Rev. A* **72**, 050306 (R) (2005).
- Rowe, M. A. *et al.* Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).
- Hong, C. K. & Mandel, L. Experimental realization of a localized one-photon state. *Phys. Rev. Lett.* **56**, 58–60 (1986).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge discussions with B. DeMarco, and partial support from the National Science Foundation, Disruptive Technologies Office, and the US Army Research Office.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to O.H. (hosten@uiuc.edu).

Quantum supercurrent transistors in carbon nanotubes

Pablo Jarillo-Herrero¹, Jorden A. van Dam¹ & Leo P. Kouwenhoven¹

Electronic transport through nanostructures is greatly affected by the presence of superconducting leads^{1–3}. If the interface between the nanostructure and the superconductors is sufficiently transparent, a dissipationless current (supercurrent) can flow through the device owing to the Josephson effect^{4,5}. A Josephson coupling, as measured by the zero-resistance supercurrent, has been obtained using tunnel barriers, superconducting constrictions, normal metals and semiconductors. The coupling mechanisms vary from tunnelling to Andreev reflection^{5–8}. The latter process has hitherto been observed only in normal-type systems with a continuous density of electronic states. Here we investigate a supercurrent flowing through a discrete density of states—that is, the quantized single particle energy states of a quantum dot⁹, or ‘artificial atom’, placed between superconducting electrodes. For this purpose, we exploit the quantum properties of finite-sized carbon nanotubes¹⁰. By means of a gate electrode, successive discrete energy states are tuned on- and off-resonance with the Fermi energy in the superconducting leads, resulting in a periodic modulation of the critical current and a non-trivial correlation between the conductance in the normal state and the supercurrent. We find, in good agreement with existing theory¹¹, that the product of the critical current and the normal state resistance becomes an oscillating function, in contrast to being constant as in previously explored regimes.

In artificial atoms, current can flow via discrete states according to the general process of resonant tunnelling, that is, ‘resonant’ when the Fermi energy in the leads is aligned with discrete energy states⁹. The maximum conductance, G , through a single spin-degenerate energy level depends on the coupling to the leads. For the case of phase-coherent tunnelling, G can reach $2e^2/h$, when charging effects are unimportant (e is the electronic charge, and h is Planck’s constant). If charging effects are significant, $G = 2e^2/h$ can still be achieved (even off-resonance) by means of the Kondo effect¹², which establishes spin coherence between the quantum dot and the leads. An entirely new situation arises in the case of superconducting leads, that is, when two superconductors are coupled via a discrete single particle state. As we show below, the conductance can reach infinity, that is, a supercurrent can flow through the quantum dot. This means that G exceeds by far the perfect conductance level of $2e^2/h$ occurring when the transmission probability reaches one. This zero resistance state is peculiar, since just a single discrete state (that can be occupied only with two spin degenerate electrons simultaneously) is available for coupling the collective macroscopic states in the leads. In contrast to previously accessible regimes, we can study Josephson coupling for on- and off-resonant tunnelling, which enables a transistor-like control of the supercurrent through the quantum dot.

The carbon nanotube (CNT) devices are fabricated by means of standard nanofabrication techniques and geometries (e-beam lithography to define customized electrodes on CNTs grown by chemical vapour deposition on top of oxidized silicon substrates¹³) with two

extra important ingredients: the choice of superconducting material and a multiple-stage filtering system to suppress electronic noise over a wide frequency range (see Fig. 1a and Supplementary Information for details).

The quantum behaviour of electrons in carbon nanotubes in good contact with metallic electrodes emerges clearly in a measurement of the differential resistance, dV/dI , versus measured source–drain voltage, V , and gate voltage, V_G , as shown in Fig. 1b for one of our devices in the normal state (I is the current through the device). The differential resistance exhibits a pattern of high and low conductance regions, typical of nanotube devices well coupled to the leads^{14,15}, with a characteristic voltage scale, $V \approx 3.5$ mV. This energy corresponds to the energy level separation between the discrete electronic states due to the finite length of the CNT, $\Delta E = \hbar v_F/2L$, where $v_F = 8.1 \times 10^5$ m s^{−1} is the Fermi velocity in the CNT, and L its length. The value obtained from this measurement, $L \approx 480$ nm, is in good agreement with the length of the nanotube segment in between the metallic electrodes, 470 nm. When the sample is cooled down below the superconducting critical temperature of the electrodes (~ 1.3 K), the electronic transport through the nanotube is strongly affected owing to the superconducting proximity effect^{1,16–19}, which can be viewed as the leakage of Cooper pairs from a superconductor into a normal metal-type material.

This proximity effect is evident from the observation of multiple Andreev reflections (MARs)²⁰ and the flow of a supercurrent through the device (Fig. 1c, d). We note that we have observed similar supercurrents in four out of seven measured metallic CNT devices with room temperature resistances below 35 k Ω (see Supplementary Information for additional data and magnetic field dependence). The most interesting feature of this supercurrent is that its maximum value (critical current, I_C) can be strongly modulated by means of a gate electrode²¹, as shown in Fig. 1d. As the CNTs are metallic, this means that the supercurrent transistor action must have a different mechanism than in conventional semiconductor structures. It is also remarkable that the gate voltage necessary to change from maximum to minimum I_C is of only ~ 50 mV, much smaller than the typical gate voltages necessary to significantly vary the charge density of semiconducting carbon nanotubes²² or nanowires with similar geometries²³.

In order to establish the origin of the modulation of I_C , it is important to characterize the sample over a larger gate voltage range. A measurement of $dV/dI(I, V_G)$ (Fig. 2b) shows a non-monotonic, quasi-periodic set of low differential resistance regions, where I_C is largest, in between regions of high dV/dI , where I_C is strongly suppressed (Fig. 2a). This pattern follows closely the low-bias pattern of Fig. 1b, with the same gate voltage spacing in between resonances, but now the vertical axis is current, instead of voltage. The correspondence between the two patterns indicates that the modulation of I_C is due to the tuning on- and off-resonance with gate voltage of the energy levels in the CNT with respect to the Fermi energy in the leads

¹Kavli Institute of Nanoscience, Delft University of Technology, PO Box 5046, 2600 GA, Delft, The Netherlands.

(as shown schematically in Fig. 2e). Such a Josephson transistor mechanism, purely due to the discrete nature of the energy levels in a nanostructure (in this case finite-sized CNTs), has not been previously observed.

Before turning to a more quantitative description, we note that the modulation of I_C is followed by a series of dV/dI peaks and dips moving up and down in the current axis. These are better seen in the high-resolution measurement shown in Fig. 2c, and reflect the MAR processes (see also Fig. 1c) taking place at the CNT–metal interfaces. MAR processes occur at voltages $V = 2\Delta_g/en$ (where Δ_g is the superconducting energy gap, and n is an integer number). The dV/dI curves in fact occur at constant voltage in this current-biased sample (see Supplementary Information). Two individual dV/dI traces are shown in Fig. 2d for the on- and off-resonance situations. In both cases dV/dI exhibits oscillations due to MAR, but the overall behaviour of dV/dI is very different. In the on-resonance case, dV/dI decreases with decreasing $|I|$ (on average) until it ‘switches’ to zero when $\sim I_C$ is reached. For the off-resonance case, dV/dI increases with decreasing $|I|$ (except at the MAR points), until $|I|$ reaches a very strongly suppressed value of I_C (barely visible in Fig. 2d, see Fig. 3). From the normal state resistance values in the on- and off-resonant cases, we can conclude that the dV/dI changes between these qualitative behaviours at values of $dV/dI \sim h/(2e^2) = 13 \text{ k}\Omega$, that is, once the resistance per channel of the CNT becomes of the order of the quantum of resistance (see also Supplementary Information).

The correlation between the critical current and the normal state resistance, R_N , is well studied in superconductor–normal metal–superconductor (SNS) structures. As a matter of fact, for short junctions in diffusive systems and ideal NS interfaces, $I_C R_N \sim \Delta_g/e$, that is, constant⁵. The situation differs when one considers a single

discrete energy level. In this case, the conductance is given by $G_N = (4e^2/h)T_{\text{BW}}$, where $T_{\text{BW}} = \Gamma_1\Gamma_2/((\epsilon_R/h)^2 + 0.25\Gamma^2)$ is the Breit–Wigner transmission probability, $\Gamma_{1,2}$ are the tunnel rates through the left/right barriers ($\Gamma = \Gamma_1 + \Gamma_2$), and ϵ_R is the energy of the resonant level relative to the Fermi energy in the leads. (Note that we have added a factor of 4 in G_N to account for the spin and orbital degeneracy of the CNT electronic states^{10,15,24,25}.) Beenakker and van Houten¹¹ have studied the lineshape for the critical current in such a system. (The general case, including finite length effects, has been studied by Galaktionov and Zaikin²⁶.) For the case of a wide resonance, $h\Gamma \gg \Delta_g$, they obtained $I_C = I_0[1 - (1 - T_{\text{BW}})^{1/2}]$, with $I_0 = 2e\Delta_g/\hbar$. Experimentally, we can vary the position of the resonant level by means of a gate voltage, $\epsilon_R \propto V_G$, as shown for the normal state conductance in Fig. 3b. From the maximum value of $G_N \approx 3.8e^2/h$, we deduce a barrier asymmetry $\Gamma_1/\Gamma_2 \approx 0.64$. We use this to fit $I_C(V_G)$ and $G_N(V_G)$ (see Fig. 3a, b; red curves). Although the functional form is in good agreement with theory, the values for Γ , $\Gamma_1 = 0.85 \text{ meV } h^{-1}$ and $\Gamma_2 = 1.36 \text{ meV } h^{-1}$, obtained from the $I_C(V_G)$ and $G_N(V_G)$ fits, respectively, differ substantially. Also the value of I_0 determined, 4.15 nA, is much smaller than the theoretical value ($I_0 = 2e\Delta_g/\hbar \approx 60 \text{ nA}$, resulting in a maximum theoretical supercurrent of 47 nA when the asymmetry in the barriers is taken into account¹¹).

Although finite length effects can yield a partial reduction of I_C (ref. 26), they cannot explain the drastic suppression (a factor ~ 15) that we measure. Such strong suppression is reminiscent of the behaviour of small, underdamped, current-biased Josephson junctions²⁷, where the electromagnetic environment leads to a measured critical current, I_{CM} , much lower than the true critical current I_C (see Supplementary Information for an estimate of the quality factor and

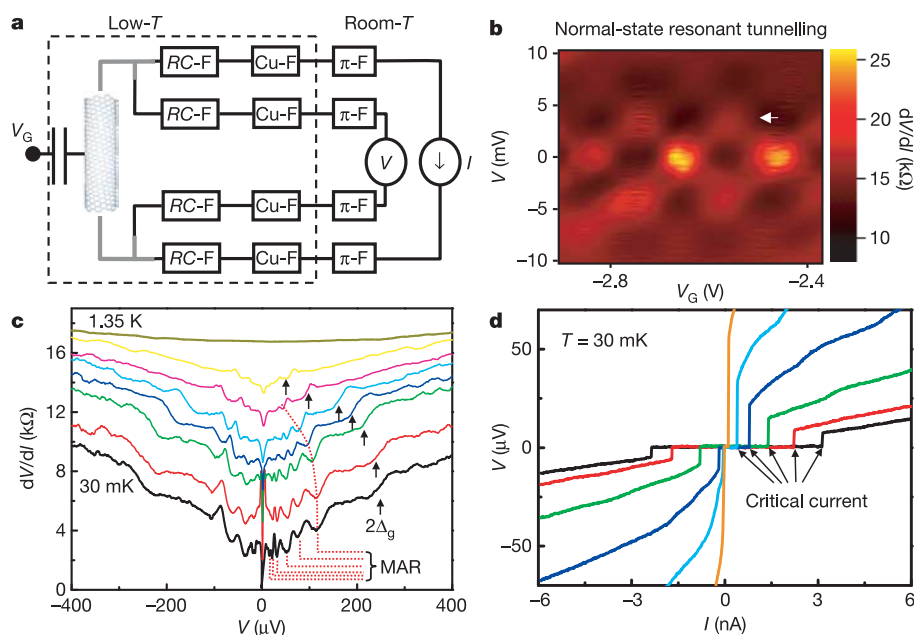


Figure 1 | Measurement scheme and basic sample characterization.

a, Diagram showing the measurement circuit. Grey represents Ti/Al electrodes (10 nm/60 nm). Titanium ensures a good electrical contact to the CNT, while aluminium becomes superconducting below $\sim 1.3 \text{ K}$, well above the base temperature of our dilution refrigerator. The CNTs are probed in a four-terminal geometry (current bias, voltage measurement). An important element is the incorporation of three sets of filters for each measurement wire: a copper-powder filter (Cu-F) for high frequency noise, π -filters for intermediate frequencies and a two-stage RC filter to suppress voltage fluctuations at low frequencies. The dashed box region indicates the low temperature part of the circuit. The rest is at room temperature. **b**, Colour-scale plot of the differential resistance, dV/dI , versus measured voltage, V , and gate voltage, V_G , at $T = 4.2 \text{ K}$. The white arrow indicates the energy

separation between discrete quantum levels in the CNT. **c**, Differential resistance versus measured source–drain voltage, V , at different temperatures (0.030, 0.47, 0.7, 0.88, 1.02, 1.18, 1.22 and 1.35 K, from bottom to top). The curves are offset for clarity (by 2 k Ω , for 0.47 and 0.7 K, and by 1 k Ω for the rest). The features present in all curves below 1.3 K are due to the induced superconducting proximity effect. The arrows indicate the superconducting gap at $V = 2\Delta_g/e \approx 250 \mu\text{V}$. The dotted lines indicate multiple Andreev reflection (MAR) processes, which manifest as dips in dV/dI . **d**, $V(I)$ characteristics at base temperature, showing the modulation of the critical current, I_C , with V_G ($V_G = -2.59, -2.578, -2.57, -2.563, -2.555$ and -2.541 V from black to orange). For currents larger than I_C the system goes into a resistive state (abrupt jump from zero to finite V).

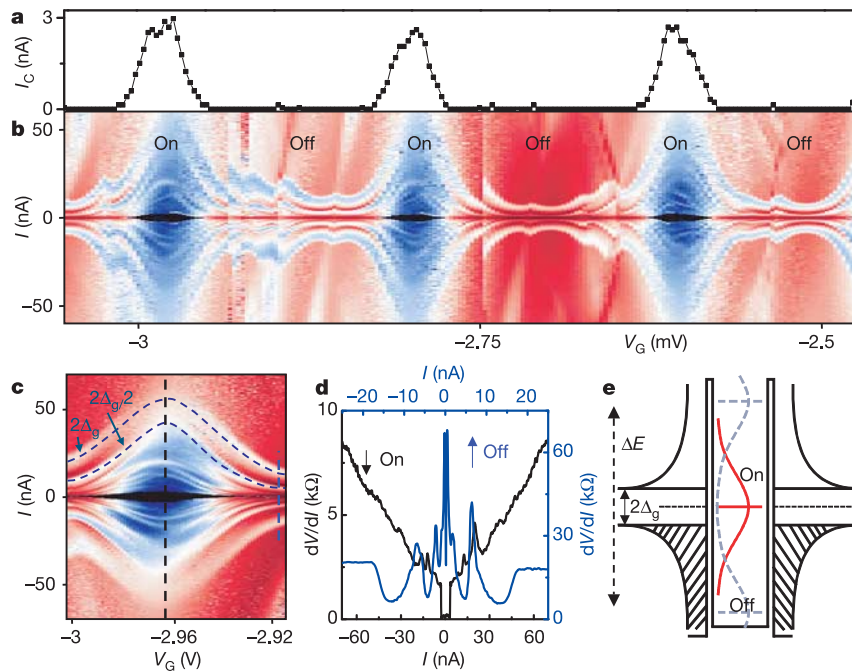


Figure 2 | Quantum supercurrent transistor. **a**, Variation of the critical current, I_C , with gate voltage, V_G , extracted from **b** (see **c** and Fig. 3a for high resolution). I_C is measured as the upper half-width of the black region around $I = 0$. **b**, Colour-scale representation (in logarithmic scale) of $dV/dI(I, V_G)$ at $T = 30$ mK (black is zero, that is, supercurrent region, and dV/dI increases from dark blue to white and red; the scale can be inferred from **d**). The differential resistance and critical current exhibit a series of quasiperiodic modulations with gate voltage as the energy levels in the CNT quantum dot are tuned on- and off-resonance with respect to the Fermi energy in the superconducting leads. The sharp vertical features are caused by random charge switches and shift the diagram horizontally. The narrow tilted features present in the off regions (for example, at $V_G \approx -2.87$ V)

occur reproducibly and are associated with Fano resonances³⁰ (see Supplementary Information). **c**, High-resolution $dV/dI(I, V_G)$ plot of the leftmost resonance region in **b**. The modulation of I_C (black central region) as well as MAR (up to several orders, the first two are highlighted by the dashed blue lines) are clearly visible. **d**, Two representative $dV/dI(I)$ curves, taken from **c** at the vertical black and blue dashed lines, illustrating the different behaviour of the differential resistance in the on- (black curve/axis) and off- (blue curve/axis) resonance case. **e**, Schematic diagram showing a strongly coupled quantum dot in between two superconducting leads. The gate voltage tunes the position of the lorentzian level from the on- (red curve) to the off- (grey dashed curve) resonance state.

suggestions for increasing I_{CM}). The dynamics of such a Josephson junction can be visualized as that of a particle moving in the so-called 'tilted washboard' potential⁵, where the driving current corresponds to the tilt in the potential. For the case of low dissipation (underdamped junctions), a small fluctuation can cause the particle to slide down the potential and go into a 'runaway' state. This occurs at a

value of I much smaller than the true I_C , and it has been shown²⁷ that the measured critical current scales as $I_{CM} \propto (I_C)^{3/2}$. In order to test the applicability of this model to CNT Josephson junctions, we have fitted $I_{CM} = I_{0M}[1 - (1 - T_{BW}(V_G, I)^{1/2})^{3/2}]$, as shown by the blue curve in Fig. 3a. We obtain a similarly low value of $I_{0M} = 4.57$ nA, and this time the value of $\hbar\Gamma_1$ obtained, 1.22 meV, is in good

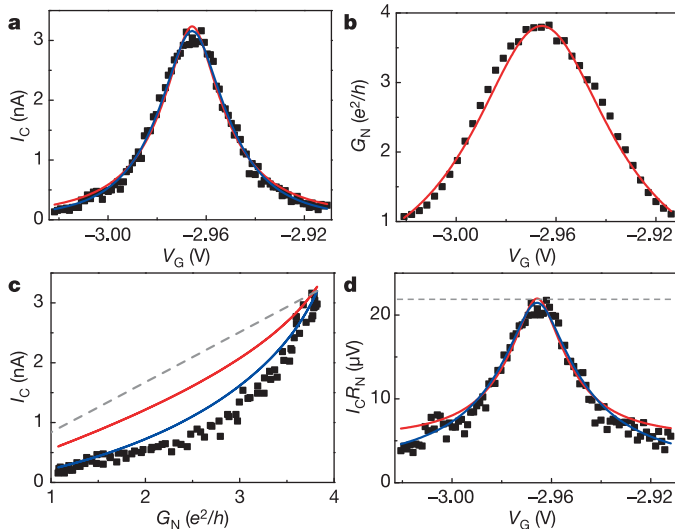


Figure 3 | Correlation between critical current and normal state conductance and modulation of the $I_C R_N$ product. In all panels, the black dots represent the experimental data points ($T = 30$ mK) and the red/blue curves are theoretical plots. **a**, Critical current, I_C , versus V_G for the resonance shown in Fig. 2c. The theoretical lines are fits to $I_C = I_0[1 - (1 - \Gamma_1\Gamma_2/((V_G - V_{GR})^2 + 0.25\Gamma^2))^{1/2}]$ (red curve) and $I_{CM} = I_{0M}[1 - (1 - \Gamma_1\Gamma_2/((V_G - V_{GR})^2 + 0.25\Gamma^2))^{1/2}]^{3/2}$ (blue), as explained in the main text. V_{GR} is the value of gate voltage on-resonance. All gate voltages and I s are converted into energies by multiplying by the gate coupling factor, $\alpha = 0.02$ meV mV^{-1} , obtained from measurements in the nonlinear regime. **b**, Conductance, G_N , as a function of V_G in the normal state ($B = 40$ mT) and the corresponding fit to $G_N = 4e^2/h(\Gamma_1\Gamma_2/((V_G - V_{GR})^2 + 0.25\Gamma^2))$. **c**, I_C - G_N correlation plot. The data show a non-trivial correlation, with a stronger decrease of I_C than expected from the theoretical curve $I_C = I_0[1 - (1 - 1/4G_N)^{1/2}]$ (red curve). The 1/4 factor simply denotes that G_N is measured in e^2/h units. The difference can be almost entirely accounted for by the influence of the electromagnetic environment, resulting in a measured $I_{CM} = I_0[1 - (1 - 1/4G_N)^{1/2}]^{3/2}$ (blue curve). An ideal SNS junction, with N a normal metal with continuous density of states, would exhibit a linear I_C - G_N correlation curve (grey dashed curve). **d**, $I_C R_N$ product versus V_G , resulting from dividing the experimental data and theory curves from **a** and **b**. The grey dashed line indicates a constant $I_C R_N$ product such as in a SNS junction.

agreement with $\hbar\Gamma_G = 1.36$ meV, resulting also in an improved fit to the data.

The importance of the coupling to the environment manifests itself more explicitly when examining the correlation between the critical current and the normal state conductance. We note that in the case of an ideal diffusive SNS junction, the correlation would yield a simple straight line. The experimental data severely deviate from such a curve (Fig. 3c). First, we consider the expected theoretical decay for the case of a discrete state (red curve) $I_C = I_0[1 - (1 - 1/4G_N)^{1/2}]$ (no fitting parameters), with the value of I_0 obtained from Fig. 3a, and G_N measured in units of e^2/h . The comparison with the predicted theoretical line shows that the measured I_C is significantly lower than expected. However, a remarkably better agreement is found when the electromagnetic environment is included, as shown by the blue curve, $I_{CM} = I_0[1 - (1 - 1/4G_N)^{1/2}]^{3/2}$, indicating the generality of the $(I_C)^{3/2}$ dependence of I_{CM} for very different types of Josephson junctions²⁷.

The predicted lineshape of I_C (even in the presence of low dissipation) implies that the $I_C R_N$ product is not constant, but instead has a maximum on-resonance. We plot in Fig. 3d the $I_C R_N$ product, which indeed exhibits a peak structure. The red and blue lines, which contain no extra fitting parameters, result from dividing the theoretical curves in Fig. 3a by the red curve in Fig. 3b, and further substantiate the results from earlier figures.

We emphasize that the above-mentioned analysis confirms the correct order of the relevant energy scales necessary for the observation of the resonant tunnelling supercurrent transistor action¹¹: ΔE (~ 3.5 meV) $> \hbar\Gamma$ (~ 1.3 meV) $\gg \Delta_g$ (~ 125 μ eV) $> U$. The last inequality is justified because signatures of Coulomb blockade effects, such as a fourfold splitting of the conductance peaks, are absent in our data, leading us to conclude that the charging energy, U , is negligible (see Supplementary Information).

We end by noting that, although both superconductivity and the Kondo effect are collective many-body phenomena, their effect on resonant tunnelling is very different¹⁸. The Kondo enhancement occurs off-resonance, while the superconducting zero-resistance state, as we have shown, is most pronounced on-resonance. In fact, we expect that the study of CNT devices with intermediate transmission, and thus, larger Coulomb interactions, will enable the observation of Kondo-enhanced supercurrents in the off-resonant case^{28,29}.

Received 19 August; accepted 21 December 2005.

- de Gennes, P. G. Boundary effects in superconductors. *Rev. Mod. Phys.* **36**, 225–237 (1964).
- Ralph, D. C., Black, C. T. & Tinkham, M. Spectroscopic measurements of discrete electronic states in single metal particles. *Phys. Rev. Lett.* **74**, 3241–3244 (1995).
- von Delft, J. & Ralph, D. C. Spectroscopy of discrete energy levels in ultrasmall metallic grains. *Phys. Rep.* **345**, 61–173 (2001).
- Josephson, B. D. Possible new effects in superconductive tunnelling. *Phys. Lett.* **1**, 251–253 (1962).
- Tinkham, M. *Introduction to Superconductivity* (McGraw-Hill, Singapore, 1996).
- Andreev, A. F. The thermal conductivity of the intermediate state in superconductors. *Sov. Phys. JETP* **19**, 1228–1231 (1964).
- Likharev, K. K. Superconducting weak links. *Rev. Mod. Phys.* **51**, 101–159 (1979).

- Blonder, G. E., Tinkham, M. & Klapwijk, T. M. Transition from metallic to tunneling regimes in superconducting micro-constrictions—excess current, charge imbalance, and super-current conversion. *Phys. Rev. B* **25**, 4515–4532 (1982).
- Sohn, L. L., Kouwenhoven, L. P. & Schön, G. (eds) *Mesoscopic Electron Transport* (Kluwer, Dordrecht, 1997).
- Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
- Beenakker, C. W. J. & van Houten, H. *Single-electron Tunneling and Mesoscopic Devices* (eds Koch, H. & Lübbig, H.) see also (<http://xxx.lanl.gov/abs/cond-mat/0111505>) (2001) 175–179 (Springer, Berlin, 1992).
- Kouwenhoven, L. & Glazman, L. Revival of the Kondo effect. *Phys. World* **14**, 33–38 (2001).
- Jarillo-Herrero, P. *et al.* Orbital Kondo effect in carbon nanotubes. *Nature* **434**, 484–488 (2005).
- Liang, W. J. *et al.* Fabry-Perot interference in a nanotube electron waveguide. *Nature* **411**, 665–669 (2001).
- Buitelaar, M. R., Bachtold, A., Nussbaumer, T., Iqbal, M. & Schonenberger, C. Multiwall carbon nanotubes as quantum dots. *Phys. Rev. Lett.* **88**, 156801 (2002).
- Kasumov, A. Y. *et al.* Supercurrents through single-walled carbon nanotubes. *Science* **284**, 1508–1511 (1999).
- Morpurgo, A. F., Kong, J., Marcus, C. M. & Dai, H. Gate-controlled superconducting proximity effect in carbon nanotubes. *Science* **286**, 263–265 (1999).
- Buitelaar, M. R., Nussbaumer, T. & Schonenberger, C. Quantum dot in the Kondo regime coupled to superconductors. *Phys. Rev. Lett.* **89**, 256801 (2002).
- Haruyama, J. *et al.* End-bonding multiwalled carbon nanotubes in alumina templates: Superconducting proximity effect. *Appl. Phys. Lett.* **84**, 4714–4716 (2004).
- Buitelaar, M. R. *et al.* Multiple Andreev reflections in a carbon nanotube quantum dot. *Phys. Rev. Lett.* **91**, 057005 (2003).
- Takayanagi, H. & Kawakami, T. Superconducting proximity effect in the native inversion layer on InAs. *Phys. Rev. Lett.* **54**, 2449–2452 (1985).
- McEuen, P. L. Single-wall carbon nanotubes. *Phys. World* **13**, 31–36 (2000).
- Doh, Y. J. *et al.* Tunable supercurrent through semiconductor nanowires. *Science* **309**, 272–275 (2005).
- Liang, W. J., Bockrath, M. & Park, H. Shell filling and exchange coupling in metallic single-walled carbon nanotubes. *Phys. Rev. Lett.* **88**, 126801 (2002).
- Sapmaz, S. *et al.* Electronic excitation spectrum of metallic carbon nanotubes. *Phys. Rev. B* **71**, 153402 (2005).
- Galaktionov, A. V. & Zaikin, A. D. Quantum interference and supercurrent in multiple-barrier proximity structures. *Phys. Rev. B* **65**, 184507 (2002).
- Joyez, P., Lafarge, P., Filipe, A., Esteve, D. & Devoret, M. H. Observation of parity-induced suppression of Josephson tunneling in the superconducting single-electron transistor. *Phys. Rev. Lett.* **72**, 2458–2461 (1994).
- Glazman, L. I. & Matveev, K. A. Resonant Josephson current through Kondo impurities in a tunnel barrier. *JETP Lett.* **49**, 659–662 (1989).
- Choi, M. S., Lee, M., Kang, K. & Belzig, W. Kondo effect and Josephson current through a quantum dot between two superconductors. *Phys. Rev. B* **70**, 020502 (2004).
- Babic, B. & Schonenberger, C. Observation of Fano resonances in single-wall carbon nanotubes. *Phys. Rev. B* **70**, 195408 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Yu. V. Nazarov, C. W. J. Beenakker, W. Belzig, S. De Franceschi and Y.-J. Doh for discussions and C. Dekker for the use of nanotube growth facilities. Financial support was obtained from the Japanese International Cooperative Research Project (ICORP) and the Dutch Fundamenteel Onderzoek der Materie (FOM).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.J.-H. (Pablo@qt.tn.tudelft.nl).

Magnetic-field-induced shape recovery by reverse phase transformation

R. Kainuma¹, Y. Imano¹, W. Ito¹, Y. Sutou², H. Morito¹, S. Okamoto³, O. Kitakami³, K. Oikawa¹, A. Fujita¹, T. Kanomata⁴ & K. Ishida¹

Large magnetic-field-induced strains¹ have been observed in Heusler alloys with a body-centred cubic ordered structure and have been explained by the rearrangement of martensite structural variants due to an external magnetic field^{1–3}. These materials have attracted considerable attention as potential magnetic actuator materials. Here we report the magnetic-field-induced shape recovery of a compressively deformed NiCoMnIn alloy. Stresses of over 100 MPa are generated in the material on the application of a magnetic field of 70 kOe; such stress levels are approximately 50 times larger than that generated in a previous ferromagnetic shape-memory alloy⁴. We observed 3 per cent deformation and almost full recovery of the original shape of the alloy. We attribute this deformation behaviour to a reverse transformation from the antiferromagnetic (or paramagnetic) martensitic to the ferromagnetic parent phase at 298 K in the Ni₄₅Co₅Mn_{36.7}In_{13.3} single crystal.

Since magnetic-field-induced strain was reported for NiMnGa (refs 1, 5), many other ferromagnetic shape-memory alloys, such as FePd (refs 2, 6, 7), FePt (ref. 8), NiCoGa (refs 9, 10), NiCoAl (refs 10–12) and NiFeGa (refs 13, 14; whose martensite phase exhibits ferromagnetism), have also been found. The origin of the extremely large magnetic-field-induced strain is explained by the rearrangement of martensite variants due to an external magnetic field, whose driving force is related to the large magnetocrystalline anisotropy energy of the martensite phase^{1–3}. According to this mechanism, because the driving force is limited by the anisotropy energy even if a large external magnetic field is applied, the output stress induced by the external magnetic field is restricted to only a few MPa (ref. 4). Magnetic-field-induced transformation is a useful method for obtaining an output stress larger than that resulting from the variant rearrangement in the martensite phase. The transformation temperature change (ΔT) induced by magnetic field change (ΔB) is approximately given by the Clausius–Clapeyron relation in the magnetic phase diagram:

$$\frac{dB}{dT} = \frac{\Delta S}{\Delta M} \quad \text{and} \quad \Delta T \approx \left(\frac{\Delta M}{\Delta S} \right) \Delta B \quad (1)$$

where T is the absolute temperature, B is the applied magnetic field, and ΔM and ΔS are the differences in magnetization and entropy between parent and martensite phases, respectively. To use magnetic-field-induced transformation, a combination of a large ΔM and a small ΔS is required for the martensitic transformation. The martensitic transformation of previous ferromagnetic shape-memory alloys is limited to the combination of ferromagnetic or paramagnetic parent and ferromagnetic martensite phases, and no large ΔT can be obtained because of the small ΔM (refs 7, 15, 16).

Very recently, our group at Tohoku University has reported some new ferromagnetic shape-memory alloys in the NiMnIn, NiMnSn and NiMnSb alloy systems¹⁷, in which the magnetization in the martensite phase is very low in comparison to that of the parent phase. Here we report that NiCoMnIn alloys, in which Co is added to NiMnIn to increase the Curie temperature, show a high level of ΔT at room temperature and that shape recovery can be obtained from magnetic-field-induced reverse transformation.

Figure 1 shows the thermomagnetization curves for the 13.4In alloy (see Methods for compositions of the three alloys we studied) at magnetic field strengths of $H = 0.5$, 20 and 70 kOe. During the first heating in a magnetic field of $H = 0.5$ kOe, the thermomagnetization curve drastically decreased between 375 K and 390 K owing to the transformation from ferromagnetic (or ferrimagnetic) to paramagnetic in the parent phase, where the Curie temperature T_C was calculated to be $T_C = 382$ K. The magnetism of the parent phase below T_C is considered to be ferromagnetic rather than ferrimagnetic, because the temperature dependence of the reciprocal susceptibility in the paramagnetic state follows the normal Curie–Weiss law. After heating to 473 K followed by furnace cooling, the

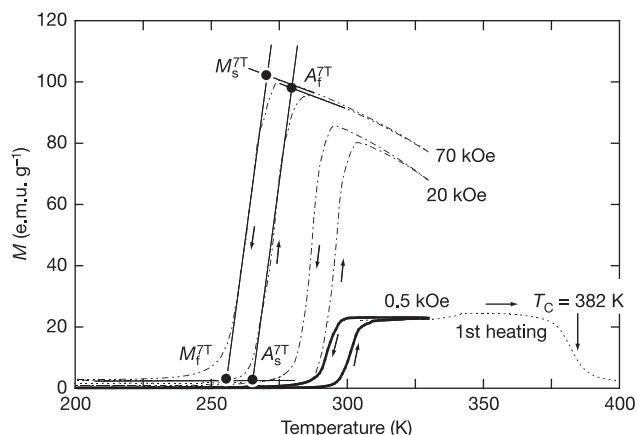


Figure 1 | Thermomagnetization curves of the Ni₄₅Co₅Mn_{36.6}In_{13.4} alloy measured in several magnetic fields by the sample extraction method. The specimen quenched from 1,173 K was first measured during heating in $H = 0.5$ kOe from room temperature to 473 K followed by furnace cooling. The magnetization change due to martensitic transformation was cyclically measured in the temperature range from 330 K to 200 K in the following order: $H = 0.5$, 20 and 70 kOe. '7T' indicates values obtained at 7 tesla ($=70$ kOe).

¹Department of Materials Science, Graduate School of Engineering, Tohoku University, 6-6-02 Aoba-yama, Sendai 980-8579, Japan. ²Tohoku University Biomedical Engineering Research Organization (TUBERO), 6-6-02 Aoba-yama, Sendai 980-8575, Japan. ³Institute of Multidisciplinary Research for Advanced Materials, Tohoku University, Katahira 2-1-1, Sendai 980-8577, Japan. ⁴Faculty of Engineering, Tohoku Gakuin University, 1-13-1 Chuo, Tagajo 980-8537, Japan.

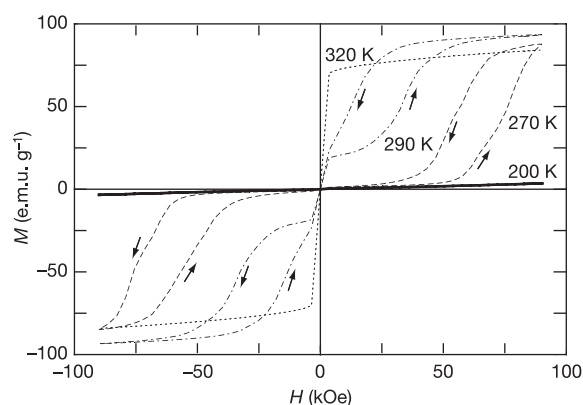


Figure 2 | Magnetization versus magnetic field curves for the $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.6}\text{In}_{13.4}$ alloy between 200 K and 320 K. The curves at 270 K and 290 K show metamagnetic transition behaviour.

magnetization was cyclically measured at temperatures between 200 K and 330 K in magnetic fields of $H = 0.5, 20$ and 70 kOe. The magnetization curve of 0.5 kOe declined suddenly at about 292 K as a result of the martensitic transformation and entered a steady region with a very low magnetization on cooling. Furthermore, the magnetization drastically increased at about 300 K as a result of the reverse transformation with a small thermal hysteresis of less than 10 K on heating.

Although the thermomagnetization curves of 20 and 70 kOe are similar to that of 0.5 kOe, all the martensitic transformation temperatures M_s, M_f, A_s and A_f decreased with increasing magnetic field, and the increase of magnetic field from 0.5 to 70 kOe resulted in a decrease in the transformation temperature of about 30 K, that is, the $\Delta T \approx 30$ K is given by $\Delta B \approx 70$ kOe in equation (1). Here, the martensitic transformation temperatures are defined as shown on the magnetization curve of 70 kOe in Fig. 1. On the basis of this experimental result and equation (1), a large $\Delta M/\Delta S$ is expected in NiCoMnIn alloys. When using $\Delta M = 100$ e.m.u. g^{-1} and $\Delta S = 27$ $\text{J kg}^{-1} \text{K}^{-1}$, which were experimentally obtained from the thermomagnetization curves in Fig. 1 and the differential scanning calorimetry (DSC) curves of the 13.4In alloy, a value of $\Delta T = 26$ K can be calculated for $\Delta B = 70$ kOe from equation (1). This value is close to the experimentally detected ΔT (~ 30 K). We note that while the magnetization of the parent phase at the M_s temperature was very large, that of the martensite phase at A_s was very low, even in a strong magnetic field of 70 kOe. However, it is not clear at present whether the magnetism of the martensite phase is antiferromagnetic or

paramagnetic (which we term non-magnetic in this paper). We confirmed by X-ray diffraction for the 13.4In powder that the parent and martensite phases have the $L2_1$ Heusler-type ordered structure where $a = 0.5978$ nm and the 14M modulated structure where $a = 0.4349$ nm, $b = 0.2811$ nm, $c = 2.9892$ nm and $\beta = 93.24^\circ$, respectively, which are comparable to those observed in the NiMnAl (ref. 18), NiMnGa (ref. 5) and NiFeGa (ref. 13) alloys.

Figure 2 shows the magnetization curves at several temperatures. While the curves at 200 and 320 K exhibit typical non-magnetic and ferromagnetic M-H behaviours, respectively, the curves at 270 and 290 K show a metamagnetic behaviour with a large hysteresis with the magnetic field inducing reverse transformation from non-magnetic martensite to the ferromagnetic parent phase. These results are in accordance with the thermomagnetization behaviour shown in Fig. 1. This magnetic behaviour is very similar to that in the FeRh alloys^{19,20} showing the first-order magnetic transition from ferromagnetism to antiferromagnetism, except that the present magnetic transition is brought about by the martensitic transformation, which also suggests the potential of NiCoMnIn alloys as magnetocaloric materials. In fact, the magnetic entropy change (ΔS_m) computed from the magnetization curves using the Maxwell relation was obtained as about 15.2 and 28.4 $\text{J kg}^{-1} \text{K}^{-1}$ at 292 K in the fields of 20 and 70 kOe, respectively. These values are comparable to those of the $\text{Gd}_5\text{Si}_2\text{Ge}_2$ (ref. 21), MnAsSb (ref. 22) and $\text{La}(\text{Fe,Si})_{13}$ (ref. 23) alloys, known as magnetocaloric materials, in which the absolute value $|\Delta S_m|$ in the field of 50 kOe is 18 $\text{J kg}^{-1} \text{K}^{-1}$ at 276 K (ref. 21), 40 $\text{J kg}^{-1} \text{K}^{-1}$ at 317 K (ref. 22) and 30 $\text{J kg}^{-1} \text{K}^{-1}$ at 184 K (ref. 23), respectively.

Here, the stress-strain characteristics of the parent and martensite phases for all the NiCoMnIn single-crystals were investigated by compressive testing, where a compressive strain of about 7% was applied at 298 K. Figure 3a to c shows the stress-strain curves of the 13.3In , 13.4In and 13.5In alloys, each with a different reverse-transformation temperature. The stress-strain characteristics at the testing temperature of $T_t = 298$ K are dependent on the A_s and A_f temperatures of the specimens, that is, when $A_f < T_t$, typical pseudoelastic behaviour with a closed loop appeared as shown in Fig. 3c, and when $A_f > T_t$, the deformed strain remained even after the stress was removed, as exhibited in Fig. 3a and b. An almost-perfect shape-memory effect on the as-deformed specimens in Fig. 3a and b was confirmed by dilatometric examination with heating to 373 K. All these results at least confirm that the NiCoMnIn alloys are shape-memory and pseudoelastic materials.

Finally, the shape recovery induced by the magnetic field was examined by a three-terminal capacitance method with the 13.3In specimen. Figure 4 shows the recovery strain induced by the magnetic field at 298 K, where a compressive pre-strain of about

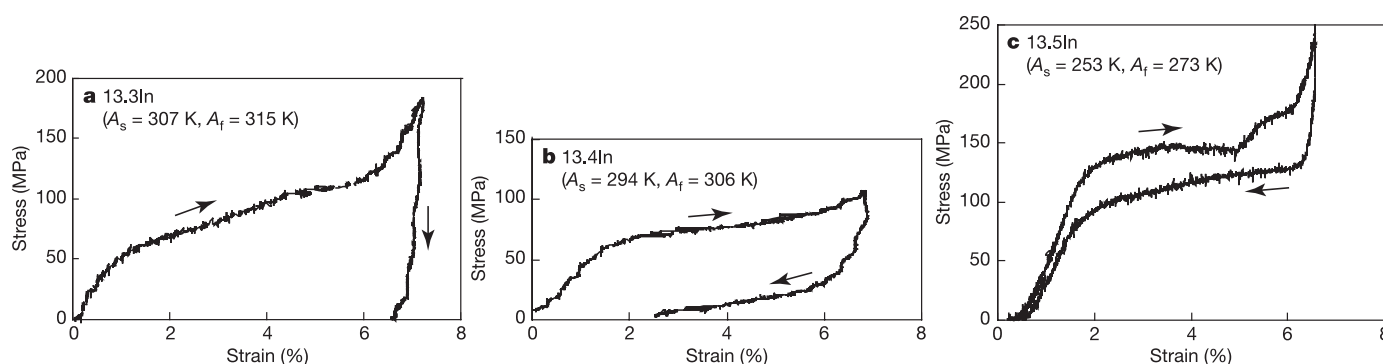


Figure 3 | Compressive stress-strain curves for the three single-crystalline alloys at $T_t = 298$ K. a, $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.7}\text{In}_{13.3}$; b, $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.6}\text{In}_{13.4}$; and c, $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.5}\text{In}_{13.5}$. The $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.5}\text{In}_{13.5}$ alloy in Fig. 3c shows almost perfect pseudoelasticity. Almost-perfect shape recovery due to

heating to 373 K after deformation was also confirmed by examination using a dilatometer for the $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.7}\text{In}_{13.3}$ and $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.6}\text{In}_{13.4}$ alloys.

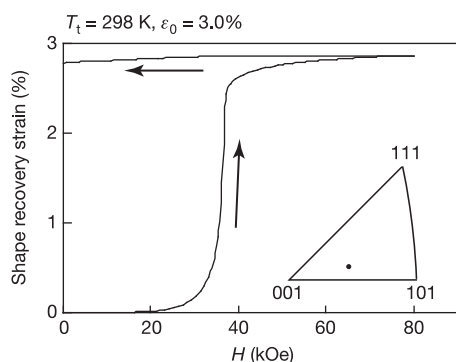


Figure 4 | Recovery strain at 298 K induced by a magnetic field for $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.7}\text{In}_{13.3}$. A compressive pre-strain of about 3% was applied to the 13.3In alloy, with the magnetic field applied in parallel to the compressive axis of the specimen and the length change parallel to the compressive axis was measured. Shape recovery is due to magnetic-field-induced reverse transformation, which is called the metamagnetic shape-memory effect.

3% was applied in the direction plotted with a filled circle in the stereographic triangle shown in Fig. 4 and the magnetic field was applied in parallel to the compressive axis of the specimen. The recovery strain started to increase at about 20 kOe, rose sharply at about 36 kOe and then gradually increased to 80 kOe with an increasing magnetic field. A recovery strain of about 2.9%, almost equal to the pre-strain of 3%, was obtained with a magnetic field of 80 kOe. However, only a very small change of length was observed from about 30 kOe when the magnetic field was removed. This behaviour induced by the magnetic field is comparable to the usual shape-memory effect due to reverse transformation induced by heating. We propose that the shape-memory effect due to magnetic-field-induced reverse transformation, that is, a metamagnetic transition, is called the 'metamagnetic shape-memory effect' and that such a shape-memory alloy is termed a 'metamagnetic shape-memory alloy'.

In contrast with previous ferromagnetic shape-memory alloys such as Ni_2MnGa and FePd , this metamagnetic shape-memory alloy system has many advantages for practical use. The most important advantage may be that the present magnetic-field-induced strain can yield a high stress proportional to the magnetic field. Critical stress of the stress-induced martensitic transformation can be calculated from the Clausius–Clapeyron relation:

$$\frac{d\sigma_C}{dT} = \frac{\Delta S}{\varepsilon V_m} \quad (2)$$

$$\Delta\sigma_C \approx \frac{\Delta S}{\varepsilon V_m} \Delta T$$

where ε is the difference of lattice strain between parent and martensite phases in a corresponding direction and V_m is the molar volume of the alloy²⁴. Using the data for the 13.4In alloy, that is, $\Delta S = 1.75 \text{ J mol}^{-1} \text{ K}^{-1}$ (determined by DSC), $\varepsilon = 0.06$ (expected as an average value for polycrystalline specimen with the 14M martensite) and $V_m = 8.05 \times 10^{-6} \text{ m}^3 \text{ mol}^{-1}$ (calculated using the lattice parameter of the parent phase), $d\sigma_C/dT$ is 3.6 MPa K^{-1} . Because the martensitic transformation temperatures decreased by about 30 K in a magnetic field of 70 kOe, the maximal obtainable critical stress of a stress-induced transformation would be $\Delta\sigma_C = 3.6 \times 30 = 108 \text{ MPa}$. This stress generated by the magnetic field is about 50 times larger than the value of about 2 MPa in the previous ferromagnetic shape-memory alloys⁴. Additionally, we expect that magnetic-field-induced strain can be obtained not only from single crystals, but also from polycrystalline specimens.

METHODS

Sample preparation. Three samples, 13.3In, 13.4In and 13.5In, with nominal compositions of $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.7}\text{In}_{13.3}$ (at.%), $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.6}\text{In}_{13.4}$ (at.%) and $\text{Ni}_{45}\text{Co}_5\text{Mn}_{36.5}\text{In}_{13.5}$ (at.%), respectively, were prepared by induction melting under an argon atmosphere. The ingots were cut into small pieces with a diamond saw and homogenized at 1,173 K for 12 to 24 h in a vacuum and then quenched in ice water. The martensitic transformation temperatures and the latent heat during the transformation were determined by DSC at heating and cooling rates of 10 K min^{-1} , and the magnetic properties were examined by the sample extraction method at heating and cooling rates of 3 K min^{-1} . The crystal structures of the parent and martensite phases were identified using the X-ray diffraction pattern of powder specimens.

Shape recovery tests. Examinations of shape recovery were performed using single-crystalline pieces with a cubic shape of about $1 \times 1 \times 2 \text{ mm}^3$, which were prepared using grain growth with arc-melting followed by annealing at 1,173 K for 1 day. The pseudoelasticity was confirmed by a compressive test and the thermally induced shape-memory effect was detected by the length change of the compressive pre-deformed specimens, using a dilatometer at a heating rate of 3 K min^{-1} . The shape recovery induced by the magnetic field was measured by a three-terminal capacitance method. The direction of the compressive stress applied to the single-crystal was confirmed by the electron back-scattering Kikuchi pattern.

Received 9 August; accepted 23 November 2005.

- Ullakko, K., Huang, J. K., Kanter, C., Kokorin, V. V. & O'Handley, R. C. Large magnetic-field-induced strains in Ni_2MnGa single crystal. *Appl. Phys. Lett.* **69**, 1966–1968 (1996).
- James, R. D. & Wuttig, M. Magnetostriction of martensite. *Phil. Mag. A* **77**, 1273–1299 (1998).
- O'Handley, R. C. Model for strain and magnetization in magnetic shape memory alloys. *J. Appl. Phys.* **83**, 3263–3270 (1998).
- Murray, S. J., Marioni, M., Allen, S. M., O'Handley, R. C. & Lograsso, T. A. 6% magnetic-field-induced strain by twin-boundary motion in ferromagnetic Ni–Mn–Ga. *Appl. Phys. Lett.* **77**, 886–888 (2000).
- Sozinov, A., Likhachev, A. A., Lanska, N. & Ullakko, K. Giant magnetic-field-induced strain in NiMnGa seven-layered martensitic phase. *Appl. Phys. Lett.* **80**, 1746–1748 (2002).
- Sakamoto, T., Fukuda, T., Kakeshita, T., Takeuchi, T. & Kishio, K. Magnetic field-induced strain in iron-based ferromagnetic shape memory alloys. *J. Appl. Phys.* **93**, 8647–8649 (2003).
- Wada, T. et al. Structural change and straining in Fe–Pd polycrystals by magnetic field. *Mater. Sci. Eng. A* **361**, 75–82 (2003).
- Kakeshita, T. et al. Giant magnetostriction in an ordered Fe_3Pt single crystal exhibiting a martensitic transformation. *Appl. Phys. Lett.* **77**, 1502–1504 (2000).
- Wuttig, M., Li, J. & Craciunescu, C. A new ferromagnetic shape memory alloy system. *Scr. Mater.* **44**, 2393–2397 (2001).
- Oikawa, K. et al. Phase equilibria and phase transformations in new B2-type ferromagnetic shape memory alloys of Co–Ni–Ga and Co–Ni–Al systems. *Mater. Trans.* **42**, 2472–2475 (2001).
- Oikawa, K. et al. Promising ferromagnetic Ni–Co–Al shape memory alloy system. *Appl. Phys. Lett.* **79**, 3290–3292 (2001).
- Morito, H. et al. Magnetocrystalline anisotropy in single-crystal Co–Ni–Al ferromagnetic shape-memory alloy. *Appl. Phys. Lett.* **81**, 1657–1659 (2002).
- Oikawa, K. et al. Magnetic and martensitic phase transitions in ferromagnetic Ni–Ga–Fe shape memory alloys. *Appl. Phys. Lett.* **81**, 5201–5203 (2002).
- Oikawa, K. et al. K. Magnetic and martensitic phase transformations in a $\text{Ni}_{54}\text{Ga}_{27}\text{Fe}_{19}$ alloy. *Mater. Trans.* **43**, 2360–2362 (2002).
- Cherechukin, A. et al. Shape memory effect due to magnetic field-induced thermoelastic martensitic transformation in polycrystalline Ni–Mn–Fe–Ga alloy. *Phys. Lett. A* **291**, 175–183 (2000).
- Sutou, Y., Kainuma, R., Ishida, K. & Taya, M. Martensitic transformation behaviour under magnetic field in Co–Ni–Al ferromagnetic shape memory alloys. *Proc. SPIE* **5053**, 159–168 (2003).
- Sutou, Y. et al. Magnetic and martensitic transformations of NiMnX (X = In, Sn, Sb) ferromagnetic shape memory alloys. *Appl. Phys. Lett.* **85**, 4358–4360 (2004).
- Kainuma, R., Gejima, F., Sutou, Y., Ohnuma, I. & Ishida, K. Ordering, martensitic and ferromagnetic transformations in Ni–Al–Mn Heusler shape memory alloys. *Mater. Trans. JIM* **41**, 943–949 (2000).
- Fallot, M. & Hocart, R. Sur l'apparition du ferromagnétisme par élévation de température dans des alliages de fer et de rhodium. *Rev. Sci.* **77**, 498–499 (1939).
- Kouvel, J. S. & Hartelius, C. C. Anomalous magnetic moments and transformations in the ordered alloy FeRh . *J. Appl. Phys.* **33**, 1343–1344 (1962).
- Pecharsky, V. K. & Gschneidner, K. A. Jr Giant magnetocaloric effect in $\text{Gd}_5(\text{Si}_2\text{Ge}_2)$. *Phys. Rev. Lett.* **78**, 4494–4497 (1997).
- Wada, H. & Tanabe, Y. Giant magnetocaloric effect of $\text{MnAs}_{1-x}\text{Sb}_x$. *Appl. Phys. Lett.* **79**, 3302–3304 (2001).

23. Fujita, A., Fujieda, S., Hasegawa, S. & Fukamichi, K. Itinerant-electron metamagnetic transition and large magnetocaloric effects in $\text{La}(\text{Fe}_{1-x}\text{Si}_x)_{13}$ compounds and their hydrides. *Phys. Rev. B* **67**, 104416 (2003).
24. Wollants, P., Debonne, M. & Roos, J. R. Thermodynamic analysis of the stress-induced martensitic-transformation in a single-crystal. *Z. Metall.* **70**, 113–117 (1979).

Acknowledgements This work was supported by a Grant-in-Aid for Core Research for Evolutional Science and Technology (CREST), Japan Science and

Technology Agency (JST) and 'Collaborative Research' in Center for Interdisciplinary Research, Tohoku University. One of the authors (R.K.) acknowledges support from the Kurata Memorial Hitachi Science and Technology Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.K. (kainuma@material.tohoku.ac.jp).

Aluminium control of argon solubility in silicate melts under pressure

M. Ali Bouhifd¹ & Andrew P. Jephcoat^{1,2}

Understanding of the crystal chemistry of the Earth's deep mantle has evolved rapidly recently with the gradual acceptance of the importance of the effect of minor elements such as aluminium on the properties of major phases such as perovskite^{1–3}. In the early Earth, during its formation and segregation into rocky mantle and iron-rich core, it is likely that silicate liquids played a large part in the transport of volatiles to or from the deep interior. The importance of aluminium on solubility mechanisms at high pressure has so far received little attention, even though aluminium has long been recognized as exerting strong control on liquid structures at ambient conditions^{4–6}. Here we present constraints on the solubility of argon in aluminosilicate melt compositions up to 25 GPa and 3,000 K, using a laser-heated diamond-anvil cell. The argon contents reach a maximum that persists to pressures as high as 17 GPa (up to 500 km deep in an early magma ocean), well above that expected on the basis of Al-free melt experiments. A distinct drop in argon solubility observed over a narrow pressure range correlates well with the expected void loss in the melt structure predicted by recent molecular dynamics simulations^{7–9}. These results provide a process for noble gas sequestration in the mantle at various depths in a cooling magma ocean. The concept of shallow partial melting as a unique process for extracting noble gases from the early Earth, thereby defining the initial atmospheric abundance, may therefore be oversimplified.

Chemical inertness and surface volatility, combined with low abundances, have made the noble gases a unique trace-elemental and isotopic system for constraining the formation and evolution of the Earth and its atmosphere^{10,11}. Over time, the Earth is assumed to have degassed, and the atmosphere formed by processes that involve escape of volatiles from silicate melts at the surface. Models of early Earth accretion include the formation of a global melt (a 'magma ocean') as a result of a giant impact with a Mars-size body¹².

In this study we investigated argon solubility in four aluminosilicate liquids at high pressures: a C1-chondrite model composition (C1-maj), C1-maj mixed with a FeNi-rich alloy, sanidine (KAlSi₃O₈), and anorthite (CaAl₂Si₂O₈) up to 25 GPa and around 3,000 K. Our main aim was to study argon solubility in silicate melts under conditions relevant to a magma-ocean stage before segregation of the Earth's mantle and core. Further, the sanidine and anorthite melts are both fully polymerized aluminosilicate glasses (with no non-bridging oxygen atoms per tetrahedral cation, NBO/T = 0), whereas C1-maj has NBO/T = 2.3, making it possible for us to investigate composition and the effect of the degree of polymerization of the melts on Ar solubility at high pressures.

The laser-heated diamond-anvil cell system used for the Ar solubility experiments has been described previously^{13,14}. Samples were mounted in the pressure chamber and loaded with dry argon using a high-pressure gas-loading technique at 200 MPa (ref. 15). Pressures were measured by the ruby fluorescence method, and

temperatures were determined spectro-radiometrically with a fit to a grey-body Planck function. Chemical analyses were made by elemental mapping across the samples (Fig. 1), avoiding any bubbles or microcrystalline phases that would distort the Ar contents of the quenched melts. The Ar solubility results presented here correspond to the Ar contents dissolved in the dense liquid phases at high pressures—there is extensive experimental evidence that densification of liquids is preserved in the glass on quenching¹⁶.

A striking feature observed during our experiments is the apparent disproportionation of the C1-maj melt at all pressures into two distinct chemical compositions (Table 1)—one an olivine composition (labelled C1-oli in Table 1) and the other a nominally Al- and Ca-enriched remnant composition (C1-rem). Both were observed from careful mapping and averaging of the recovered glass samples. The regions of distinct composition were found from electron microprobe averages over a micrometre scale and not clearly delineated at any single boundary (see Supplementary Information).

The experimental argon solubility of each melt as a function of pressure is plotted in Fig. 2a–c and all show good agreement with Ar

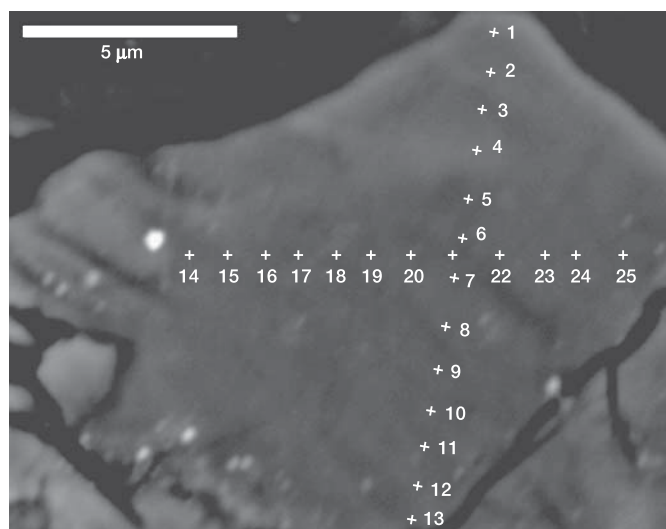


Figure 1 | Typical back-scattered electron image of a polished sample that was recovered from 10 GPa and 2,500 K with C1-maj as the starting material. Typically samples were quenched by switching off the laser power, removed from the diamond-anvil cell and mounted into epoxy for polishing. The samples have been examined by micro-Raman spectroscopy to confirm the amorphous state of the quenched run products and by electron microprobe and scanning electron microscopy for the chemical analyses. The sample surface shows only a glassy texture, as confirmed by Raman spectra. The numbers shown in the figure are the order of the analyses (in a composition scan) providing evidence for the existence of two distinct chemical compositions in the melt (see Supplementary Information).

¹University of Oxford, Department of Earth Sciences, Parks Road, Oxford OX1 3PR, UK. ²Diamond Light Source Ltd, Chilton, Didcot OX11 0QX, UK.

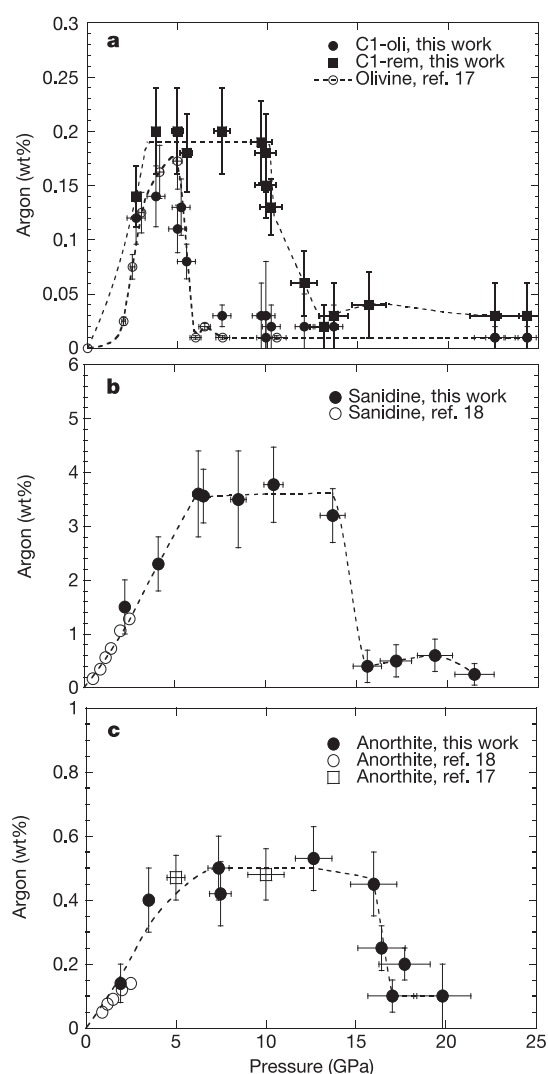


Figure 2 | Ar contents as a function of pressure. **a**, For C1-rem and C1-oli melts up to 25 GPa. The Ar solubility data for olivine up to 10.5 GPa are those of ref. 17. **b**, For sanidine melt up to 22 GPa. The data at low pressure (up to 2.5 GPa) are those of ref. 18. **c**, For anorthite melt up to 20 GPa. The Ar solubility at 5 and 10 GPa are data reported by ref. 19, and the data at low pressure (up to 2.5 GPa) are those of ref. 18. The reported Ar contents represent an average of 20 analyses in the areas with no bubbles or microcrystalline phases that would distort the chemical analysis of argon. The dashed lines are a guide to the eye. Error bars correspond to 2σ of the chemical analysis, and to 10% of the nominal pressures reported¹⁴.

solubility trends in experiments at lower pressures^{17–19}. For all melts, the amount of dissolved argon increases with pressure up to about 5 GPa, to reach a maximum of 0.2, 0.5 and 3.5 wt% for C1-rem, anorthite and sanidine melts, respectively. These low-pressure differences are consistent with the well-known dependence of rare-gas

solubility on the silicate melt composition below 2.5 GPa (refs 18, 20). At higher pressures, the Ar content remained constant in the range 5–17 GPa, before a marked drop of the Ar solubility for the three melts: C1-rem at around 10 GPa, sanidine near 15 GPa and anorthite at about 17 GPa (Figs 2a–c).

To simulate Earth's core formation under conditions of segregation from a deep magma ocean, we performed experiments with a starting mix of C1-maj and a Fe–Ni alloy. The run products from this mix indicate that the oxygen fugacity, f_{O_2} , is about two orders of magnitude below the iron–wüstite buffer and is in accord with core-formation models in which metallic liquid equilibrates with molten silicate under reducing conditions²¹. The maximum Ar content we measured in C1-rem melt is ~ 0.2 wt%, corresponding to $8.9 \times 10^{-5} \text{ cm}^3 \text{ g}^{-1} \text{ bar}^{-1}$ at standard temperature and pressure (with a C1-rem glass density of 2.75 g cm^{-3}), and is of the same order of magnitude as the value derived from basaltic melts equilibrated with Ar at low partial pressures^{22–24}. This value indicates that the levels of argon obtained in the experiments are applicable to planetary differentiation processes. Further, comparison of the argon solubility in the C1-oli melt only with the existing data for olivine melt up to 12 GPa matches the characteristic drop of Ar solubility observed near 5 GPa (ref. 17). This agreement directly supports the apparent separation we observe of the C1-maj melt into an olivine component, and a residue. These C1-melt observations indicate that the minor elements Ca and Al exhibit a 'melt–melt' partitioning under pressure that mirrors the differences in compositions of the major solid mineral assemblages (garnet and olivine) of the upper mantle.

Taken together, all these results indicate that a discontinuous drop of Ar solubility at high pressure should be a universal behaviour of silicate melts, but—critically—with the onset pressure depending on the chemical composition of the melt and, by inference, its local structure. Previous work based on the behaviour of Ar in olivine and SiO_2 had predicted that all mantle melts would exhibit a low Ar solubility above a low threshold pressure of 3–5 GPa (ref. 17). But our experiments show clearly that the onset pressure for reduced argon solubility is critically dependent on the Al/(Al + Si) ratio, and correlates well with an increase in the ratio Al/(Al + Si) of the melts (Fig. 3).

One implication of these results is that different mechanisms of compression of Al-free and Al-bearing silicate melts must take place at high pressure and must correlate with the variation in the onset transition pressures to low-argon solubilities: evidence that aluminosilicate melts respond differently under pressure is provided by nuclear magnetic resonance studies of quenched glasses that show the network aluminium atoms to be more susceptible to the formation of high-coordinated species than silicon^{4,5}. More importantly, the mechanism of formation of five-coordinated atoms is restricted to aluminium sites in the network, suggesting that not only the degree of polymerization, but also the ratio of (Al/Al + Si) strongly influences the behaviour of the aluminosilicate liquid structures at high pressures^{4,5}.

Molecular dynamics simulations of molten anorthite indicate that above 20 GPa the activation volumes for all melt species drop below

Table 1 | Chemical composition of the run products (element wt%)

	C1-maj	C1-rem†	C1-oli†	C1-rem‡	C1-oli‡	Sanidine	Anorthite
O*	44.8 ± 0.1	45.7 ± 1.4	44.5 ± 1.4	45.0 ± 1.4	41.0 ± 2.0	46.1 ± 2.0	45.7 ± 2.0
Mg	21.1 ± 0.5	21.4 ± 1.9	31.2 ± 1.4	21.8 ± 1.5	31.3 ± 1.9		
Al	1.8 ± 0.2	3.3 ± 1.0	0.5 ± 0.4	3.2 ± 0.9	0.9 ± 0.5	9.4 ± 0.3	19.3 ± 1.0
Si	23.2 ± 0.5	22.4 ± 1.5	19.2 ± 0.7	20.5 ± 1.5	19.4 ± 0.8	29.3 ± 1.4	19.9 ± 0.7
Ca	2.4 ± 0.2	2.0 ± 0.3	0.5 ± 0.2	2.4 ± 0.5	0.9 ± 0.5		13.9 ± 1.0
Fe	6.6 ± 0.5	5.0 ± 1.4	4.1 ± 1.5	6.4 ± 1.0	6.3 ± 1.7	1.4 ± 0.6	1.1 ± 0.4
Ni				0.6 ± 0.4	0.4 ± 0.2		

* Oxygen is calculated by stoichiometry.

† The starting material is a pure C1-maj glass.

‡ The starting material is a mix of C1-maj glass (65.5 wt%), Fe (24.2 wt%) and Ni (10.3 wt%) metal powders (ref. 14).

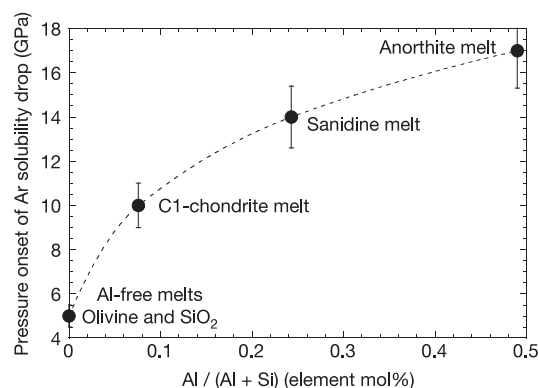


Figure 3 | Pressure onset of the Ar solubility drop in silicate melts at high pressure versus Al/(Al + Si). For Al-free melts, such as olivine and SiO₂ melts, with Al/(Al + Si) = 0 (refs 17 and 19), the Ar solubility drops at 5 GPa, whereas this decrease is observed near 10, 14 and 17 GPa for the C1-rem, sanidine and anorthite melts (with Al/(Al + Si) = 0.076, 0.243 and 0.49, respectively). Error bars correspond to 10% of the reported pressures¹⁴.

the partial molar volume of argon⁷—we observe the solubility step near 17 GPa—and hence a first-order interpretation of the Ar drop data is that they represent a closing of a characteristic void space in melts at high pressures and temperatures. Noble gases have long been conceptually accommodated in ‘holes’ in silicate melts²⁰. With increasing pressure, the fraction of highly coordinated Si and Al gradually increases (from [4]Si or [4]Al to [5]Si and [6]Si or [5]Al and [6]Al, respectively)²⁵. This general (qualitative) mechanism leads to a continuous densification of aluminosilicate melts under pressure, and our results show clearly that a minimum porosity (corresponding to a maximum pressure) is shown below which noble gases (in this case, argon) cannot be accommodated in the melt structure.

We now address possible controls on the compositional dependence of Ar solubility in silicate melts at high pressures. As stated above, for pressures below 5 GPa (Fig. 4), Ar solubility increases linearly with increasing SiO₂ content (mol%) independently of the network modifiers or formers present in the melt, suggesting that Ar is most soluble in SiO₂-rich liquids^{17–20,26,27}. In Fig. 4 we also plot argon content in several melts versus mol% SiO₂ from a range of literature studies at or near 8 GPa: at this pressure, the Ar content is no longer a simple function of the SiO₂ content of the melt. In aluminium-free melts, the NBO/T ratio was taken as a good predictor of argon solubility²⁷. The wide variation of Ar content (from 0.5 to 3.5 wt%) observed for the fully-polymerised sanidine and anorthite melts here further suggest that NBO/T is not in fact likely to be a good predictor for Ar solubility in natural melts.

A number of consequences of these trends in solubility are relevant to the ‘planetary processing’ of volatile components: Below 5 GPa (corresponding to depths less than 150 km in the mantle), argon may dissolve in silicate melts when they form by partial melting in the Earth’s interior—that is, argon maintains its geochemically assumed role as normally ‘incompatible’ with crystalline phases. In a study up to 8 GPa, the partition coefficient for argon between clinopyroxene and its melt was found to be low ($\sim 10^{-3}$ – 10^{-4}), implying that argon would be (substantially) expressed into the melt and also ultimately degassed²⁸. But, the high-pressure step decrease of Ar solubility in aluminosilicate melts we observe in the range of 10–17 GPa simply implies that the change of behaviour from high to moderate compatibility with the crystalline phase is postponed: there are no apparent contradictions in trends of argon solubility at high pressures²⁶ because the ‘transition’ in solubility simply occurs at higher pressures with aluminium present.

At depths for pressures exceeding 10 GPa (the minimum for C1-rem), the presence of melt or solid no longer strongly determines where the argon is retained in the Earth, opening up the possibility

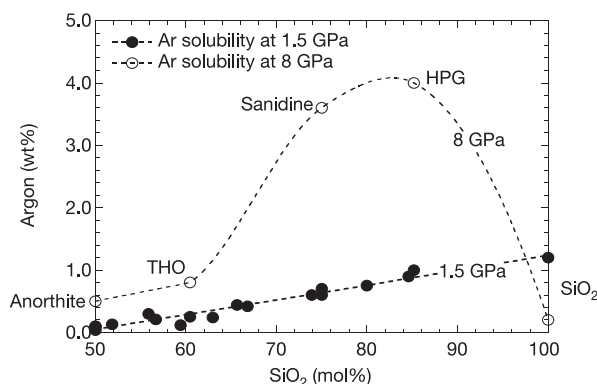


Figure 4 | Ar solubility as a function of SiO₂ content (mol%) at 1.5 and 8 GPa. The data are taken from refs 17–19 and 26 and this work. (HPG and THO are an haplogranitic and tholeiitic melts, respectively.)

once again that argon (and other noble gases at different pressures) can remain in the mantle—and there are geochemical arguments that much of the Ar budget for example must remain in a mantle reservoir²⁹. Cooling of a magma ocean would have led to varying melt composition with depth as solid phases are extracted that would control the distribution and ultimate degassing of argon within an extended depth range at least to the top of the transition zone.

Of course, our experiments are performed under argon-saturated conditions in the diamond-anvil cell where the absolute argon pressure is much higher than any likely early Earth partial pressure (although it is possible that an early dense atmosphere existed), but nonetheless the results do indicate that it would be possible for primordial and radiogenic argon to be taken deep into the Earth. In much the same way that aluminium has radically changed the phase transformations at the transition zone¹, the compression mechanism of aluminous perovskite², and the oxidation state chemistry of perovskite in the lower mantle³, it is probably also responsible for significant changes in silicate liquid chemistry within the deep early Earth.

METHODS

Starting materials. The glasses were prepared separately from oxide and carbonate mixes through repeated cycles of grinding and fusion at around 1,873 K and rapidly cooled to obtain the glasses. All the starting materials were powders with a grain size of particles less than 5 μ m. To be able to heat sanidine and anorthite glasses with the YAG laser, roughly 10 wt% of either Fe metal or Pt powders were added to separate mixes. Each composition was mixed and ground under ethanol, and finally dried at around 150 °C for at least 24 h. The C1-maj composition is (49.6 wt% SiO₂, 35.1 wt% MgO, 8.5 wt% FeO, 3.5 wt% Al₂O₃ and 3.3 wt% CaO). The proportions of SiO₂, Al₂O₃, MgO and CaO were chosen to be that of model C1-chondrite. The mix of C1-chondrite with an FeNi-alloy consisted of 66 wt% of C1-maj, 24 wt% Fe metal and 10 wt% Ni metal).

High-pressure set-up. For our LHDAC (laser-heated diamond anvil cell) Ar solubility experiments, we used diamond anvils with 500- μ m culets, and stainless-steel gaskets pre-indented to a thickness of around 50 μ m and drilled to a diameter of 150 μ m. Samples were mounted in the pressure chamber and dry-argon-loaded with a high-pressure gas-loading technique at 200 MPa (ref. 15). The pressures given in this study are those measured by the ruby fluorescence method and the hydrostatic pressure scale. Such pressures may differ from the real pressure of the sample during heating, because thermal pressure may be generated in the laser spot. In our configuration it is impossible to quantify the effects of the thermal pressure, but an increase of a few GPa on top of the nominal pressures reported was deduced from the LHDAC studies of olivines and periclase. The temperatures were determined spectroradiometrically with a fit to a grey-body Planck function (the temperatures in our experiments ranged between 2,500 and 3,000 K for all pressures). During melting of silicate under pressure, the surrounding solid argon melts by conductive heating from the heated sample, allowing dissolution of Ar into the silicate melts. To reduce the temperature gradient across the samples, we used a relatively broad, defocused beam (hot-spot size around 20 μ m, with central temperature gradients of less than 20 K μ m⁻¹). With this temperature gradient,

the maximum difference between the temperatures at the centre and the edge of the laser spot is 200 K, which was taken as a conservative uncertainty on the temperatures. During heating, when the melt is detected optically, the laser power was adjusted to give a temperature higher by about 200° in comparison with the estimated melting temperatures.

Analytical techniques. The chemical compositions of the run products were analysed by electron microprobe (JEOL JXA-8800R with four wavelength-dispersive spectrometers) with an accelerating voltage of 15 kV and a beam current of 20 nA, as well by scanning electron microscope (JEOL JSM-840A) which is equipped with an Oxford Instruments Isis 300 energy-dispersive analytical system (the accelerating voltage used was 20 kV, with a beam current of 6 nA). Fe, Ni and Co metals, orthoclase, jadeite, wollastonite, periclase and corundum were used as standards for both instruments. For Ar analyses, the standard used was an anorthite glass from this work containing 1.5 wt% Ar.

Argon analysis. The reported Ar contents in this paper represent an average of 20 analyses in the areas with no bubbles or microcrystalline phases that would distort the chemical analysis of argon. Analyses repeated more than ten times on the same area showed no difference in the Ar content. Furthermore, we checked the Ar content of several quenched samples polished to different depths, especially those with high and low Ar contents for the same starting material. The Ar contents remained constant up to 20 µm from the raw surface, showing the reliability of the Ar analyses and the homogenous mixing of argon with the silicate melts, which rules out any exsolution during sample quench (see Supplementary Information).

Received 15 July 2005; accepted 12 January 2006.

- Wood, B. J. & Rubie, D. C. The effect of alumina on phase transformations at the 660-kilometer discontinuity from Fe-Mg partitioning experiments. *Science* **273**, 1522–1524 (1996).
- Brodholt, J. P. Pressure-induced changes in the compression mechanism of aluminous perovskite in the Earth's mantle. *Nature* **407**, 620–622 (2000).
- McCammon, C. The paradox of mantle redox. *Science* **308**, 807–808 (2005).
- Yarger, J. et al. Al coordination changes in high-pressure aluminosilicate liquids. *Science* **270**, 1964–1967 (1995).
- Poe, B. T. et al. Silicon and oxygen self-diffusivities in silicate liquids measured to 15 gigapascals and 2800 Kelvin. *Science* **276**, 1245–1248 (1997).
- Mysen, B. O. & Richet, P. *Silicate Glasses and Melts, Properties and Structure* 231–290 (Elsevier, Amsterdam, 2005).
- Nevins, D. & Spera, F. J. Molecular dynamics simulations of molten $\text{CaAl}_2\text{Si}_2\text{O}_8$: Dependence of structure and properties on pressure. *Am. Mineral.* **83**, 1220–1230 (1998).
- Rustad, J. R., Yuen, D. A. & Spera, F. J. Molecular dynamics of liquid SiO_2 under high pressure. *Phys. Rev. A* **42**, 2081–2089 (1990).
- Trave, A., Tangney, P., Scandolo, S., Pasquarello, A. & Car, R. Pressure-induced structural changes in liquid SiO_2 from ab-initio simulations. *Phys. Rev. Lett.* **89**, 245504 (2002).
- Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* Edn 2 (Cambridge Univ. Press, Cambridge, 2002).
- Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92**, 2–79 (1991).
- Tonks, W. B. & Melosh, H. J. Magma ocean formation due to giant impacts. *J. Geophys. Res.* **98**, 5319–5333 (1993).
- Jephcoat, A. P. & Besedin, S. in *Properties of Earth and Planetary Materials at High Pressure and Temperature*. *Geophys. Monogr.* **101**, 287 (1998).
- Bouhifd, M. A. & Jephcoat, A. P. The effect of pressure on partitioning of Ni and Co between silicate and iron-rich metal liquids: a diamond-anvil cell study. *Earth Planet. Sci. Lett.* **209**, 245–255 (2003).
- Jephcoat, A. P., Mao, H. K. & Bell, P. M. in *Hydrothermal Experimental Techniques* 469–506 (Wiley-Interscience, New York, 1987).
- Poe, B. T., Romano, C., Zotov, N., Cibin, G. & Marcelli, A. Compression mechanisms in aluminosilicate melts: Raman and XANES spectroscopy of glasses quenched from pressures up to 10 GPa. *Chem. Geol.* **174**, 21–31 (2001).
- Chamorro-Perez, E., Gillet, P., Jambon, A., Badro, J. & McMillan, P. Low argon solubility in silicate melts at high pressure. *Nature* **393**, 352–355 (1998).
- White, B. S., Brearly, M. & Montana, A. Solubility of argon in silicate liquids at high pressures. *Am. Mineral.* **74**, 513–529 (1989).
- Chamorro-Perez, E., Gillet, P. & Jambon, A. Argon solubility in silicate melts at very high pressures. Experimental set-up and preliminary results for silica and anorthite melts. *Earth Planet. Sci. Lett.* **145**, 97–107 (1996).
- Carroll, M. R. & Stolper, E. M. Noble gas solubilities in silicate melts and glasses: New experimental results for argon and the relationship between solubility and ionic porosity. *Geochim. Cosmochim. Acta* **57**, 5039–5051 (1993).
- Thibault, Y. & Walter, M. J. The influence of pressure and temperature on the metal-silicate partition coefficients of nickel and cobalt in a model C1 chondrite and implications for metal segregation in a deep magma ocean. *Geochim. Cosmochim. Acta* **59**, 991–1002 (1995).
- Hiyagon, H. & Ozima, M. Partitioning of noble-gases between olivine and basalt melt. *Geochim. Cosmochim. Acta* **50**, 2045–2057 (1986).
- Jambon, A., Weber, H. & Braun, O. Solubility of He, Ne, Ar, Kr and Xe in a basalt melt in the range 1250–1600 °C. Geochemical implications. *Geochim. Cosmochim. Acta* **50**, 401–408 (1986).
- Lux, G. The behaviour of noble gases in silicate liquids: solution, diffusion, bubbles, and surface effects, with applications to natural samples. *Geochim. Cosmochim. Acta* **51**, 1549–1560 (1987).
- Allwardt, J. R., Stebbins, J. F., Burkhard, S. & Frost, D. J. in *Frontiers in High-Pressure Research* (eds Chen, J., Wang, Y., Duffy, T. S., Shen, G. & Dobrzynetskaia, L. F.) 211–240, (2005).
- Schmidt, B. C. & Keppler, H. Experimental evidence for high noble gas solubilities in silicate melts under mantle pressures. *Earth Planet. Sci. Lett.* **195**, 277–290 (2002).
- Shibata, T., Takahashi, E. & Matsuda, J. Solubility of neon, argon, krypton, and xenon in binary and ternary silicate systems: a new view on noble gas solubility. *Geochim. Cosmochim. Acta* **62**, 1241–1253 (1998).
- Brooker, R. A. et al. The “zero charge” partitioning behaviour of noble gases during mantle melting. *Nature* **423**, 738–741 (2003).
- Allègre, C. J., Hofmann, A. & O’Nions, R. K. The argon constraints on mantle structure. *Geophys. Res. Lett.* **23**, 3555–3557 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Richet and D. Porcelli for their comments, P. Richet for donating the anorthite sample material, A. K. Klepe and N. R. Charnley for their help during this work. M.A.B. acknowledges the earlier support of a Marie Curie EU Fellowship. This work was supported by NERC Grants to A.P.J.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A.B. (alib@earth.ox.ac.uk) or A.P.J. (andrew@earth.ox.ac.uk).

Tunicates and not cephalochordates are the closest living relatives of vertebrates

Frédéric Delsuc^{1†}, Henner Brinkmann¹, Daniel Chourrout² & Hervé Philippe¹

Tunicates or urochordates (appendicularians, salps and sea squirts), cephalochordates (lancelets) and vertebrates (including lamprey and hagfish) constitute the three extant groups of chordate animals. Traditionally, cephalochordates are considered as the closest living relatives of vertebrates, with tunicates representing the earliest chordate lineage^{1,2}. This view is mainly justified by overall morphological similarities and an apparently increased complexity in cephalochordates and vertebrates relative to tunicates². Despite their critical importance for understanding the origins of vertebrates³, phylogenetic studies of chordate relationships have provided equivocal results^{4–7}. Taking advantage of the genome sequencing of the appendicularian *Oikopleura dioica*, we assembled a phylogenomic data set of 146 nuclear genes (33,800 unambiguously aligned amino acids) from 14 deuterostomes and 24 other slowly evolving species as an outgroup. Here we show that phylogenetic analyses of this data set provide compelling evidence that tunicates, and not cephalochordates, represent the closest living relatives of vertebrates. Chordate monophyly remains uncertain because cephalochordates, albeit with a non-significant statistical support, surprisingly grouped with echinoderms, a hypothesis that needs to be tested with additional data. This new phylogenetic scheme prompts a reappraisal of both morphological and palaeontological data and has important implications for the interpretation of developmental and genomic studies in which tunicates and cephalochordates are used as model animals.

The introduction of molecular data into classical systematics has already tested several evolutionary hypotheses through the analysis of individual genes such as ribosomal RNA (rRNA). However, phylogenies reconstructed from a single gene or a small number of genes are hampered by stochastic effects limiting the statistical significance of the results. The genomic era is now providing the opportunity for phylogenetics to resolve a number of outstanding evolutionary questions through an increase of resolving power⁸. This applies to the origin and early evolution of vertebrates, a fundamental evolutionary question that has been revived by recent advances in molecular and developmental biology as well as new fossil discoveries³.

Our understanding of these events has to be considered in the context of chordate phylogeny, for which the traditional textbook view considers cephalochordates as the closest living relatives of vertebrates (a group named Euchordata), to the exclusion of the morphologically more distinct tunicates². Although almost universally accepted, this classical picture is supported by only a limited number of morphological features that are far from being unambiguous. For example, the presence of metameric segmentation¹ used to link cephalochordates and vertebrates might in fact be considered as an ancestral feature of deuterostomes⁹. The classical view (Euchordata) has also found some support in molecular studies

of rRNA genes⁵. However, a competing hypothesis grouping tunicates and vertebrates into a clade named Olfactores¹⁰ was recovered in cladistic analyses of combined rRNA and morphology⁴ and suggested by the structure of cadherin genes¹¹. However, the statistical significance of these apparently conflicting results was limited by the relatively few characters considered.

Recently, two multigene studies based on nuclear proteins have provided some support for Olfactores^{6,7}. However, the extremely limited chordate species sampling considered in these studies meant that no firm conclusions were possible, given the potentially deleterious effects of poor taxon sampling on phylogenetic inference⁸. We have therefore extended the 146-gene data set of ref. 7 from 4 to 13 chordates, including one cephalochordate, four tunicates, and eight vertebrates, with the notable inclusion of the early-branching agnaths (hagfish and lamprey). Within tunicates, the incorporation of *O. dioica* is particularly important because it belongs to appendicularians (or larvaceans), which are morphologically and molecularly very divergent from the ascidians previously included.

Phylogenetic analyses of this multigene data set using maximum parsimony (MP), maximum likelihood (ML) and bayesian inference all converged to the same topology (Fig. 1). The statistical support was maximal in bayesian analyses, where all nodes received a posterior probability of 1.0. All non-controversial groups (choanoflagellates, cnidarians, molluscs, arthropods, tunicates and vertebrates) were recovered with strong bootstrap support (bootstrap percentage BP_{MP-ML} > 95%), as were also metazoans, bilaterians, protostomes and lophotrochozoans (Fig. 1). Weak statistical support (BP_{MP-ML} < 72%) was only observed for some relationships within insects and bivalves. Within vertebrates, our results strongly support the controversial monophyly of cyclostomes (lamprey and hagfish)². Also, both MP and ML provided reasonable support for the monophyly of deuterostomes (BP_{MP-ML} = 87–93%). However, within deuterostomes, chordates appeared not to be monophyletic, because cephalochordates grouped with echinoderms, albeit with moderate ML bootstrap support (BP_{MP-ML} = 97–89%). By contrast, whereas MP moderately supported the grouping of tunicates and vertebrates (90%), the more accurate ML method¹² provided unambiguous bootstrap support (100%) for Olfactores (Fig. 1).

To test further the stability of phylogenetic relationships within deuterostomes, we evaluated in a likelihood framework the 15 rooted topologies corresponding to all possibilities of connecting the four major groups under study (echinoderms, cephalochordates, tunicates and vertebrates). Thirteen alternatives to the ML topology were significantly rejected at the 5% confidence level by all statistical tests (Table 1). Only the topology where chordates are monophyletic, with cephalochordates joining the tunicates plus vertebrate clade, was not rejected (Table 1). The traditional hypothesis of euchordate monophyly was ranked only fourth in terms of log-likelihood, after the

¹Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada. ²Sars Centre for Marine Molecular Biology, Bergen High Technology Centre, University of Bergen, Thormøhlensgaten 55, 5008 Bergen, Norway. [†]Present address: Laboratoire de Paléontologie, Phylogénie et Paléobiologie, Institut des Sciences de l'Évolution, UMR 5554-CNRS, Université Montpellier II, France.

alternative in which cephalochordates emerge before echinoderms. These two topologies appeared significantly worse than the ML tree of Fig. 1, even for the conservative Shimodaira–Hasegawa test¹³.

Our results therefore indicate a strong phylogenetic affinity between tunicates and vertebrates to the exclusion of cephalochordates. However, obtaining high statistical support for a given topology does not necessarily indicate that the phylogenetic inference is correct. Indeed, the phylogenetic analysis of large-scale data sets requires particular attention to potential systematic biases associated, for instance, with differences in evolutionary rates among species, compositional biases and heterotachy⁸. In particular, a long-branch attraction (LBA) artefact¹⁴ may potentially occur since tunicates include fast (*Ciona* spp.) and very fast (*O. dioica*) evolving species (Fig. 1). A high evolutionary rate of tunicate genes was already noticed in rRNA genes⁵ and in complete mitochondrial genomes¹⁵. Our results confirm these observations for a large number of nuclear genes. As fast evolutionary rates are also often associated with compositional bias or with heterotachy, it is a necessary first step to exclude the possibility that the observed grouping of tunicates with vertebrates results from a tree reconstruction artefact.

The most obvious potential artefact, an LBA, predicts that the fast-evolving tunicates would be attracted towards the outgroup, and not by the slowly evolving vertebrates. This would produce a topology compatible with the classical hypothesis of chordate evolution where the slow-evolving cephalochordates and vertebrates group together. This prediction is perfectly congruent with the lower support for Olfactores observed with MP (90%), a method known to be more sensitive to LBA than probabilistic methods¹⁴. Indeed, when *O. dioica*

was used as the single representative of tunicates, MP unambiguously supported (BP = 100) an aberrant position for this group which emerged before cnidarians, disrupting the monophyly of bilaterians (Supplementary Fig. S1). By contrast, the less sensitive ML method recovered Olfactores, albeit with decreased bootstrap support (BP = 84) (Supplementary Fig. S2). Therefore, despite its extreme evolutionary rate, *O. dioica* retained enough phylogenetic signal for its position to be recovered with ML. This demonstrates that LBA is not responsible for the inferred grouping of tunicates and vertebrates, and represents a strong argument in favour of the authenticity of Olfactores. In addition, neither compositional bias nor heterotachy significantly influenced phylogenomic inference with our data set (see Supplementary Information). In fact, the compositional effect would act against Olfactores, because vertebrates and the amphioxus share similar amino acid compositions (Supplementary Fig. S3). In conclusion, the strongly supported monophyly of Olfactores cannot be explained by any kind of identifiable systematic bias (LBA, compositional bias, and heterotachy) and therefore constitutes the best current hypothesis for chordate phylogeny.

The monophyly of deuterostomes remained moderately supported in our phylogenomic analyses (Fig. 1). Also, the monophyly of chordates is not found in the ML tree, but is the only alternative not significantly rejected by likelihood-based statistical tests (Table 1). A unique origin of chordates and their distinctive features such as notochord and hollow nerve cord cannot be excluded. Our results nevertheless favoured the intriguing possibility of a sister-group relationship between cephalochordates and echinoderms that seems robust to analyses aimed at avoiding compositional bias and

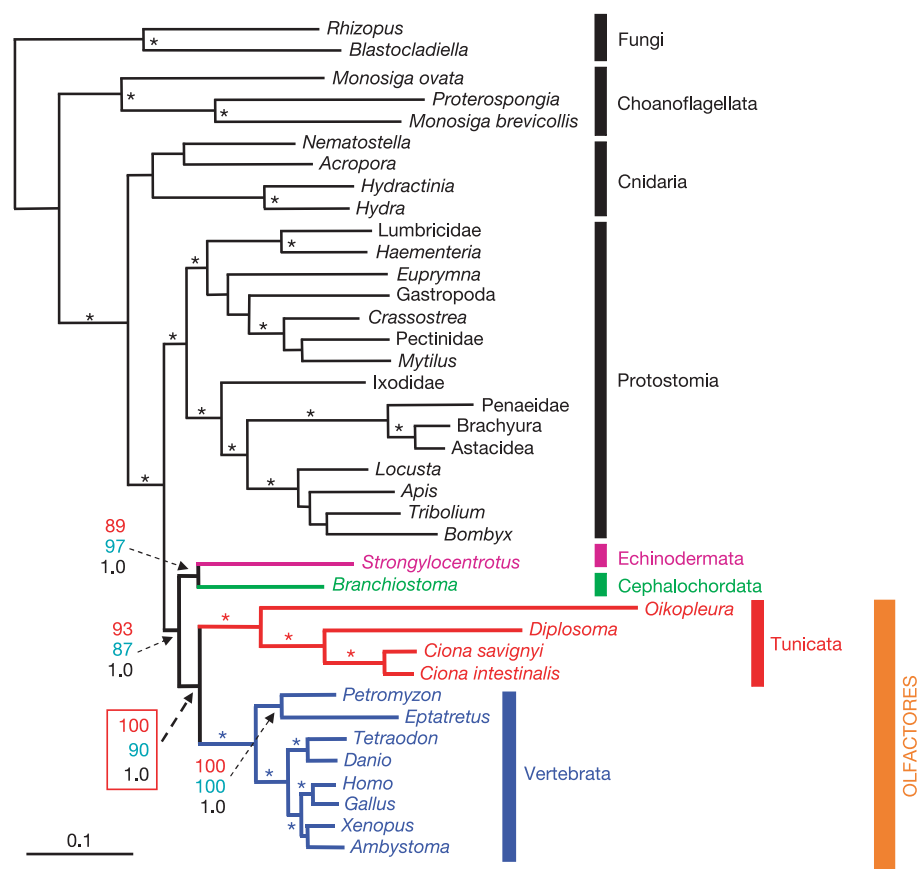


Figure 1 | Phylogenetic analyses of genomic data strongly support the grouping of tunicates and vertebrates into Olfactores. ML tree obtained from the analysis of 33,800 aligned amino acid positions under a WAG substitution matrix plus a four-category gamma rate correction ($\alpha = 0.5$) using two independent reconstruction algorithms (see Supplementary Information). Weighted maximum parsimony and bayesian inference using

the same WAG + F + Γ_4 model and WAG + F + Γ_4 plus covarion model also retrieved the same topology (see Supplementary Information). Bootstrap proportions obtained after 100 ML (red) and 1,000 MP replicates (blue), as well as bayesian posterior probabilities (black) are shown for selected branches. A star indicates that all three values are maximal (100%, 100% and 1.0). Scale bar indicates number of changes per site.

heterotachy (see Supplementary Information). Such a relationship has also been inferred from mitochondrial genomes¹⁶, but it lacks significant statistical support in both nuclear and mitochondrial analyses.

Although seemingly heretic, the grouping of echinoderms and cephalochordates constitutes an interesting working hypothesis. A similar situation was encountered a few years ago for the recently established sister-group relationship of echinoderms and hemichordates (Ambulacraria). The Ambulacraria hypothesis led to a re-evaluation of morphological character evolution, with the presence of pharyngeal slits being interpreted as an ancestral feature of deuterostomes⁹. Similarly, a close relationship between echinoderms and cephalochordates would imply that a dorsal nerve chord was already present in the last common deuterostome ancestor and subsequently evolved into derived nervous systems in both hemichordates and echinoderms. Such a scenario seems a priori possible given the difficulties encountered in polarizing morpho-anatomical characters in both extant⁹ and fossil¹⁷ deuterostomes. However, a definitive conclusion will only be achieved through the phylogenetic analysis of more genes combined with an increased taxon sampling including the enigmatic xenoturbellidans, hemichordates, and a greater diversity of echinoderms. Nonetheless, the strong support obtained for Olfactores will probably not be affected, as these additional taxa are considered to be on the echinoderm side of the deuterostome tree^{5,18}. This prediction is supported by the observation that removing the sea-urchin from our data set has almost no effect (Supplementary Fig. S4).

Despite this remaining uncertainty, our new phylogenetic hypothesis implies a serious re-evaluation of fundamental aspects of deuterostome evolution. The nature of their last common ancestor has been most extensively addressed from the paleontological point of view³. However, extant deuterostome lineages are morphologically so distinct that possible stem-group representatives found in the fossil record are difficult to recognize¹⁷. A sister-group relationship of tunicates and vertebrates to the exclusion of cephalochordates is compatible with the controversial calcichordate theory of chordate origins proposed by Jefferies¹⁹. However, it does not mean that this evolutionary scenario based on the functional reconstruction of

unusual fossils with calcite skeletons (cornutes and mitrates) and their interpretation as stem-group chordates¹⁹ is necessarily true. In fact, Jefferies¹⁰ coined the name Olfactores on the basis of the presence of a putatively homologous olfactory apparatus in fossils that were proposed to be precursors of tunicates and vertebrates. However, the phylogenetic position of cornutes and mitrates is still much debated, with the majority advocating echinoderm affinities for these controversial fossils^{20,21}. At any rate, the present molecular evidence for a monophyletic group of tunicates and vertebrates might help to polarize morphological characters of basal deuterostome fossils, thereby leading to a better understanding of early deuterostome evolution.

Our results also prompt a reinterpretation of morphological data in deuterostome phylogeny. In particular, a close proximity between tunicates and vertebrates suggests that the presence of metameric segmentation classically used to unify cephalochordates and vertebrates might be considered as an ancestral feature that underwent a secondary reduction in tunicates⁹. More generally, this new phylogenetic picture is in agreement with an alternative hypothesis for chordate evolution based on a recent homology analysis of morphological structures in hemichordates and chordates²². This unorthodox view proposes that cephalochordates have retained many ancestral characters that have been secondarily lost in the morphologically more derived tunicates, and reveals 13 putative synapomorphies uniting tunicates and vertebrates to the exclusion of cephalochordates²². The monophyly of Olfactores invalidates the traditional textbook representation of chordate, and even deuterostome, evolution as a steady increase towards complexity culminating in the highly specialized brain of vertebrates⁹. This anthropocentric interpretation is perhaps best reflected by the terms 'Euchordata' (that is, 'true chordates') or 'chordates with a brain', which are used to designate the grouping of cephalochordates and vertebrates². Tunicates should therefore no longer be considered as 'primitive' but rather as derived chordates with highly specialized lifestyles and developmental modes.

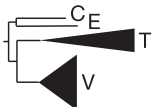
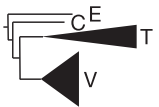
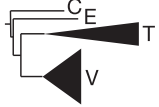
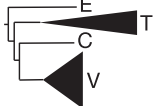
From the developmental point of view, our phylogenetic results help to understand the origin of the major evolutionary novelty constituted by the neural crest. This vertebrate innovation can be traced back to the origins of the chordate lineage, because 'latent homologues' of neural crest cells have been identified in both cephalochordates and tunicates²³. However, evidence for migratory neural crest cells has so far been reported only in tunicates²⁴, whereas their existence is still unproved in amphioxus. In light of the Olfactores hypothesis, these migratory cells may well have evolved in the last common ancestor of tunicates and vertebrates, after the divergence from cephalochordates, with these evolutionary precursor cells later giving birth to the neural crest along the vertebrate lineage²⁴.

The newly proposed deuterostome phylogeny strengthens the view that tunicates and cephalochordates represent complementary models for studying the origin of the vertebrate developmental program. Indeed, tunicates are phylogenetically closer to vertebrates but are morphologically and molecularly highly derived with a trend towards genomic simplification^{25,26}, whereas the more distantly related cephalochordates might have retained more ancestral characters²⁷. The comparative analysis of available tunicate and vertebrate genomes with the upcoming amphioxus and sea-urchin genome sequences will be particularly valuable for understanding the evolution of new gene systems and structures involved in early vertebrate development.

METHODS

Data assembly. We built upon a phylogenomic data set consisting of 146 nuclear genes previously assembled to study animal phylogeny⁷. This data set was updated using the same protocol (see Supplementary Information) with new sequences publicly available from the Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/>) and the EST Database (<http://www.ncbi.nlm.nih.gov/dbEST/>) of

Table 1 | Results of likelihood-based tests of alternative topologies within deuterostomes

Trees	−ln[L]	Δln[L]	AU	SH	RELL BP
	554,914.8	Best	0.947	1.000	0.938
	554,967.2	52.4	0.071	0.415	0.061
	555,051.5	136.7	0.000*	0.019*	0.000*
	555,066.4	151.6	0.004*	0.007*	0.002*

Results computed with a concatenated WAG + F + T₄ model are given for the top four ranking topologies; the other 11 alternative topologies were rejected by all tests with $P < 0.001$. L, likelihood value; AU, Approximately Unbiased test; SH, Shimodaira–Hasegawa test; RELL BP, Resampling of Estimated Log-Likelihood Bootstrap Percentage. *Statistically significant at the 5% level. E, echinoderms; C, cephalochordates; T, tunicates; V, vertebrates.

GenBank at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). Sequences from the appendicularian were generated by the *Oikopleura dioica* genome project (http://www.genoscope.cns.fr/externe/English/Projets/Projet_HG/HG.html).

As previously demonstrated^{7,8}, taxon sampling has a major impact on phylogenomic studies. As an outgroup to the 14 available deuterostomes, we therefore selected the slowest-evolving taxa among available protostomes and fungi to reduce the potential impact of long-branch attraction¹⁴. Furthermore, we also incorporated all available cnidarians and choanoflagellates, allowing us efficiently to break the long branch leading to the distantly related fungal outgroup.

Phylogenetic analyses. Multiple methods using different optimality criteria and algorithms were used to analyse our phylogenomic data set. Weighted MP heuristic searches were conducted using PAUP²⁸ with ten random additions of species and Tree Bisection and Reconnection branch swapping. MP bootstrap percentages were obtained after 1,000 replications using the same heuristic search strategy. Because of the computational difficulties involved in conducting ML searches for such a large data set⁷, ML analyses were conducted with different algorithms (see Supplementary Information for details). ML bootstrap percentages were obtained after 100 replications. Bayesian phylogenetic inferences were also conducted using parallel computing (see Supplementary Information for details).

Likelihood-based tests of alternative topologies were calculated using CONSEL¹³. ML branch lengths of alternative topologies were first inferred assuming a concatenated WAG + F + Γ_4 model using TREE-PUZZLE²⁹, site-wise log-likelihood values were then computed with CODEML³⁰ and the *P* values of the different likelihood-based tests were finally calculated with CONSEL.

Received 9 August; accepted 19 October 2005.

- Schaeffer, B. Deuterostome monophyly and phylogeny. *Evol. Biol.* **21**, 179–235 (1987).
- Rowe, T. in *Assembling the Tree of Life* (eds Cracraft, J. & Donoghue, M. J.) 384–409 (Oxford Univ. Press, Oxford, 2004).
- Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, London, 1996).
- Zrzavy, J., Mihulka, S., Kepka, P., Bezdek, A. & Tietz, D. Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics* **14**, 249–285 (1998).
- Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J. & Mallatt, J. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol. Biol. Evol.* **19**, 762–776 (2002).
- Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.* **22**, 2275–2284 (2005).
- Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol. Biol. Evol.* **22**, 1246–1253 (2005).
- Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nature Rev. Genet.* **6**, 361–375 (2005).
- Gee, H. in *Major Events in Early Vertebrate Evolution: Palaeontology, Phylogeny, Genetics, and Development* (ed. Ahlberg, P. E.) 1–14 (Taylor and Francis, London, 2001).
- Jefferies, R. P. S. in *Biological Asymmetry and Handedness* (eds Bock, G. R. & Marsh, J.) 94–127 (Wiley, Chichester, 1991).
- Oda, H., Akiyama-Oda, Y. & Zhang, S. Two classic cadherin-related molecules with no cadherin extracellular repeats in the cephalochordate amphioxus: distinct adhesive specificities and possible involvement in the development of multicell-layered structures. *J. Cell Sci.* **117**, 2757–2767 (2004).
- Felsenstein, J. *Inferring Phylogenies* (Sinauer, Sunderland, Massachusetts, 2004).
- Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
- Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410 (1978).
- Yokobori, S., Oshima, T. & Wada, H. Complete nucleotide sequence of the mitochondrial genome of *Doliolum nationalis* with implications for evolution of urochordates. *Mol. Phylogenet. Evol.* **34**, 273–283 (2005).
- Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguna, J. & Boore, J. L. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol. Phylogenet. Evol.* **33**, 321–332 (2004).
- Conway Morris, S. The Cambrian “explosion”: slow-fuse or megatonnage? *Proc. Natl Acad. Sci. USA* **97**, 4426–4429 (2000).
- Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. & Telford, M. J. *Xenoturbella* is a deuterostome that eats molluscs. *Nature* **424**, 925–928 (2003).
- Jefferies, R. P. S. *The Ancestry of the Vertebrates* (Cambridge Univ. Press, London, 1986).
- Peterson, K. J. A phylogenetic test of the calcichordate scenario. *Lethaia* **28**, 25–38 (1995).
- Jefferies, R. P. S. A defence of the calcichordates. *Lethaia* **30**, 1–10 (1997).
- Ruppert, E. E. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* **83**, 8–23 (2005).
- Stone, J. R. & Hall, B. K. Latent homologies for the neural crest as an evolutionary novelty. *Evol. Dev.* **6**, 123–129 (2004).
- Jeffery, W. R., Strickler, A. G. & Yamamoto, Y. Migratory neural crest-like cells form body pigmentation in a urochordate embryo. *Nature* **431**, 696–699 (2004).
- Seo, H. C. et al. Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431**, 67–71 (2004).
- Edwards, R. B. et al. Remodelling of the homeobox gene complement in the tunicate *Oikopleura dioica*. *Curr. Biol.* **15**, R12–R13 (2005).
- Holland, L. Z., Laudet, V. & Schubert, M. The chordate amphioxus: an emerging model organism for developmental biology. *Cell. Mol. Life Sci.* **61**, 2290–2308 (2004).
- Swofford, D. L. PAUP*: *Phylogenetic Analyses Using Parsimony and Other Methods* (Sinauer, Sunderland, Massachusetts, 2000).
- Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504 (2002).
- Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Conway Morris, R. P. S. Jefferies, W. R. Jeffery and J. Mallatt for suggestions, and N. Lartillot and N. Rodrigue for critical readings of early versions of the manuscript. *Oikopleura* genome data were generated at Génoscope Evry (France) with material and co-funding from the Sars International Centre. We are grateful to P. Wincker and the Génoscope team. We gratefully acknowledge the financial support provided by Génome Québec, the Canadian Research Chair and the Université de Montréal, and the Réseau Québécois de Calcul de Haute Performance for computational resources.

Author Contributions H.P. conceived the study. D.C. contributed sequence data from the *Oikopleura* genome project. F.D., H.B. and H.P. assembled the data set and performed phylogenetic analyses. F.D. wrote the first draft of the manuscript and all authors contributed to the writing of its final version.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.P. (herve.philippe@umontreal.ca).

Parasitic plants indirectly regulate below-ground properties in grassland ecosystems

Richard D. Bardgett¹, Roger S. Smith², Robert S. Shiel³, Simon Peacock², Janet M. Simkin², Helen Quirk¹ & Phil J. Hobbs⁴

Parasitic plants are one of the most ubiquitous groups of generalist parasites in both natural and managed ecosystems, with over 3,000 known species worldwide^{1–3}. Although much is known about how parasitic plants influence host performance^{1–4}, their role as drivers of community- and ecosystem-level properties remains largely unexplored⁵. Parasitic plants have the potential to influence directly the productivity and structure of plant communities because they cause harm to particular host plants, indirectly increasing the competitive status of non-host species^{6–10}. Such parasite-driven above-ground effects might also have important indirect consequences through altering the quantity and quality of resources that enter soil, thereby affecting the activity of decomposer organisms^{3,11–13}. Here we show in model grassland communities that the parasitic plant *Rhinanthus minor*, which occurs widely throughout Europe and North America¹⁴, has strong direct effects on above-ground community properties, increasing plant diversity and reducing productivity. We also show that these direct effects of *R. minor* on the plant community have marked indirect effects on below-ground properties, ultimately increasing rates of nitrogen cycling. Our study provides evidence that parasitic plants act as a major driver of both above-ground and below-ground properties of grassland ecosystems.

Rhinanthus minor is a facultative root hemiparasite that is commonly associated with low to medium fertility grasslands¹⁵ (see Supplementary Fig. 1) and is known to infect fast-growing grasses as a preferred host, thereby reducing their competitive dominance¹⁶. As a consequence, the use of *R. minor* has been proposed as a management tool for restoring botanical diversity to agriculturally improved grassland^{8,17}. We set up mesocosms of mixed grassland communities on soils of varying residual fertility, based primarily on soil phosphorus (P) availability, ranging from low, to low-medium, to medium fertility (see Methods and Supplementary Table 1). Grassland communities were constructed by sowing into field soil a number of grass (*Agrostis capillaris*, *Alopecurus pratensis*, *Briza media*, *Lolium perenne*, *Phleum bertolonii*, *Poa trivialis*) and forb (*Geranium sylvaticum*, *Lotus corniculatus*, *Ranunculus acris*, *Ranunculus bulbosus*, *Ranunculus repens*) species typical of traditionally managed grassland¹⁸. To these, three *R. minor* treatments were applied to represent a range of natural field densities (that is, none, medium density at 30 plants m⁻², and high density at 60 plants m⁻²)^{8,14}, along with three manure treatments (that is, none, 12 tonnes ha⁻¹ alternate years, and 12 tonnes ha⁻¹ yr⁻¹). This factorial design of treatments allowed us to test, in soils of varying residual fertility, the influence of *R. minor* on community properties relative to normal farming practice of applying manures¹⁹.

Total above-ground biomass (excluding *R. minor*) measured three years after treatments had started was reduced significantly by high

densities of *R. minor* ($F_{2,54} = 4.72$, $P = 0.013$) (Fig. 1a). Furthermore, there were no significant interactions with residual soil fertility or manure application (Table 1), indicating that the suppressive effect of *R. minor* on above-ground biomass occurred irrespective of

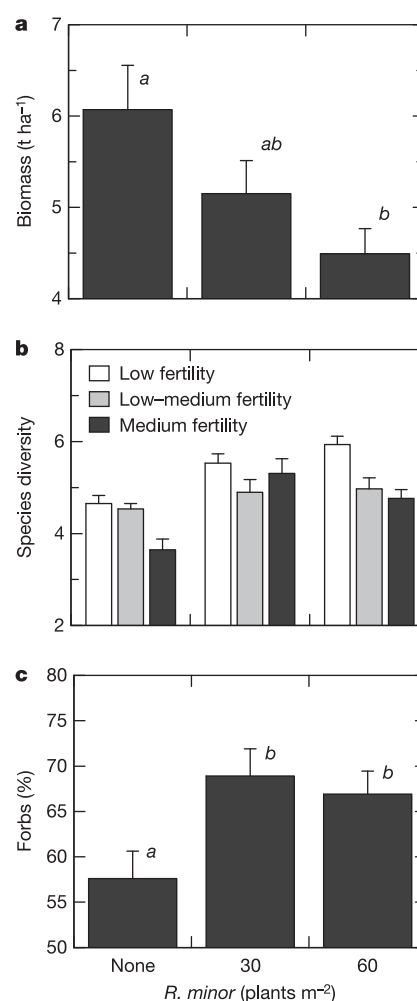


Figure 1 | Mean ($\pm$ s.e.m.) values for plant community parameters as affected by the presence of *R. minor* after three years. **a**, Biomass of above-ground plant material, excluding *R. minor*. **b**, Plant diversity, excluding *R. minor*, calculated using the inverse Simpson's index. **c**, Proportion of forbs, excluding *R. minor*, within the plant community. Values with different letters are significantly different at the $P < 0.05$ level.

¹Institute of Environmental and Natural Sciences, Soil and Ecosystem Ecology Laboratory, Lancaster University, Lancaster LA1 4YQ, UK. ²School of Biology and Psychology, ³School of Agriculture, Food and Rural Development, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK. ⁴Institute of Grassland and Environmental Research, North Wyke Research Station, Okehampton, Devon EX20 2SB, UK.

Table 1 | Above-ground community properties after three years of experimental treatments

Source	Species composition		Above-ground biomass		Diversity	
	Per cent*	F (P-value)	Per cent*	F (P-value)	Per cent*	F (P-value)
Main treatment effects						
FYM addition	3.1	1.74 (NS)	1.3	0.62 (NS)	5.0	3.73 (0.031)
<i>R. minor</i> addition	19.2	10.89 (0.002)	9.9	4.72 (0.013)	26.0	19.52 (<0.001)
Residual fertility	11.5	6.51 (0.002)	4.0	1.94 (NS)	14.6	10.89 (<0.001)
Two-way interactions						
<i>R. minor</i> × residual fertility	6.0	1.76 (0.012)	1.1	0.25 (NS)	8.2	3.06 (0.024)
<i>R. minor</i> × FYM addition	3.0	0.85 (NS)	7.6	1.82 (NS)	6.1	2.29 (NS)
FYM × residual fertility	2.3	0.62 (NS)	0.8	0.19 (NS)	0.7	0.27 (NS)
Three-way interaction						
FYM × <i>R. minor</i> × soil residual fertility	5.3	–	18.8	2.25 (0.038)	3.4	0.63 (NS)
Residual and block variance	49.6	–	56.5	–	36.0	–

* Per cent variation attributable to treatment.

FYM, farmyard manure; NS, $P > 0.05$.

variation in these factors. Although above-ground biomass was greatest ($F_{2,54} = 47.78$, $P < 0.001$) on the most fertile soils and in pots receiving manure ($F_{2,54} = 19.33$, $P < 0.001$) after one year, these responses were no longer detected after three years (Supplementary Table 2). Despite this, shoot tissue P concentration (a proxy for soil P availability) for the grass *L. perenne* and forbs *G. sylvaticum* and *R. acris* was lowest in the low fertility soil (Supplementary Table 3). Also, root colonization of these three plant species with arbuscular mycorrhizal fungi, which is typically highest when P is deficient²⁰, was consistently greater in the low fertility soil (Supplementary Table 3). Together, these findings indicate that initial differences in soil fertility, in terms of P availability, continued throughout the experiment.

Associated with a decline in above-ground productivity was a significant increase in plant diversity ($F_{2,54} = 19.52$, $P < 0.001$) in the presence of *R. minor* (Fig. 1b). Plant species diversity was also affected significantly by residual soil fertility, being greatest in the low fertility soil ($F_{2,54} = 10.89$, $P < 0.001$) (Fig. 1b). There was also a significant interaction ($F_{4,54} = 3.06$, $P = 0.024$) between the effects on plant diversity of residual soil fertility and *R. minor*: plant diversity was always lowest at zero *R. minor* for any given soil fertility, and was always largest in the low residual fertility soil at any given *R. minor* level, but the significance of differences between these treatments were not consistent (Fig. 1b). The application of manure resulted in a significant increase in plant diversity ($F_{2,54} = 3.73$, $P = 0.031$). Analysis of individual plant responses to *R. minor* revealed that its positive effect on plant diversity was due to a marked suppression of the competitively dominant grass *L. perenne* ($F_{2,54} = 41.3$, $P < 0.001$), and an increase in the abundance of the forbs *G. sylvaticum* ($F_{2,54} = 17.61$, $P < 0.001$) and *R. acris* ($F_{2,54} = 35.76$, $P < 0.001$) (Table 2). As a consequence, *R. minor* markedly increased ($F_{2,54} = 4.57$, $P = 0.015$) the proportion of forbs within the community (Fig. 1c). This is consistent with the prediction that *R. minor* will increase grassland diversity through suppression of competitively dominant grasses, which this species preferentially infects¹⁶. The abundance of these grass and forb species was unaffected by the application of manure, but this treatment did significantly decrease the proportion of forbs within the community ($F_{2,54} = 3.91$, $P = 0.026$).

The significance of *R. minor* as a driver of plant community composition, relative to the other factors of residual soil fertility and manure application, is highlighted by the fact that most variance in measures of plant community biomass and composition was attributable to its presence; only a small proportion of variance was attributed to residual fertility and manure, and interactions between experimental factors (Table 1). These data suggest, therefore, that within low-to-moderately productive mesotrophic grasslands, where *R. minor* typically occurs¹⁵, parasite-driven effects can be equal to, or greater than, those of other factors such as residual soil fertility and manure application.

Producer and decomposer subsystems function in tandem to maintain ecosystem functioning^{21–23}. Therefore, we expected above-ground responses to *R. minor* to be associated with changes in decomposer processes and organisms. To test this, we measured soil microbial community responses to *R. minor* and other treatments using phospholipid fatty acid analysis (PLFA) (see Methods). Total PLFA was unaffected by experimental treatments, indicating that the size of the microbial community was unresponsive. However, we found that the structure of the microbial community changed. In particular, the ratio of fungal-to-bacterial fatty acids was lowest in the most fertile soil ($F_{2,54} = 8.98$, $P < 0.001$), which is consistent with the notion that fungi are relatively more abundant and functionally important than bacteria under less fertile conditions^{22,23} (Fig. 2a). The fungal-to-bacterial biomass ratio measure was also reduced by the presence of *R. minor*, albeit with marginal significance ($F_{2,54} = 2.53$, $P = 0.09$) (Fig. 2b), suggesting that changes in the plant community resulting from *R. minor* infection set a trend in motion towards increasing bacterial dominance in the microbial community.

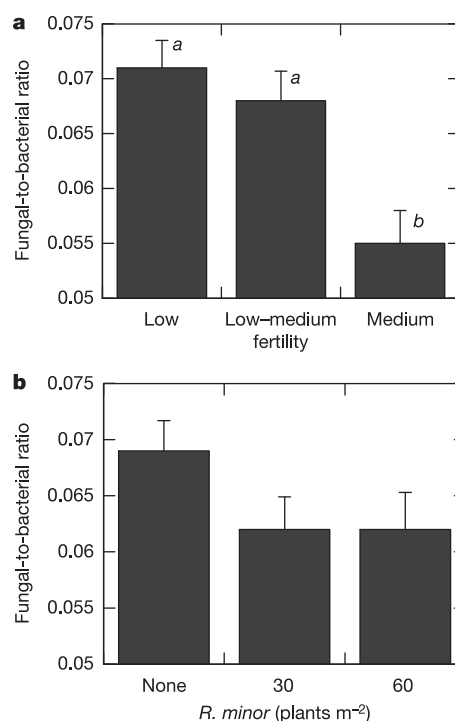


Figure 2 | Mean ($\pm$ s.e.m.) values for fungal-to-bacterial biomass ratio, calculated using PLFA, after three years. a, Influence of soil residual fertility. b, Influence of *R. minor*. Values with different letters are significantly different at the $P < 0.05$ level.

Table 2 | Significant treatment effects of *R. minor* on the biomass of two forbs and a grass after three years

Species	<i>Rhinanthus minor</i> population		
	None	30 plants m ⁻²	60 plants m ⁻²
<i>Geranium sylvaticum</i>	77.0 ^a	313.8 ^b	286.6 ^b
<i>Ranunculus acris</i>	327.5 ^a	896.2 ^b	870.6 ^b
<i>Lolium perenne</i>	1,872.0 ^a	664.3 ^b	592.3 ^b
Other species	3,793.5	3,555.7	2,970.4

Values are in kg ha⁻¹, and different letters indicate that values are significantly different at the $P < 0.05$ level. Other species refers to the total biomass of all other species that showed no response to the *R. minor* treatment plus the biomass of *R. minor*. *G. sylvaticum* and *R. acris* are forbs; *L. perenne* is a grass.

Increases in the abundance of bacteria relative to fungi in the soil microbial community are typically associated with an increase in rates of nutrient cycling^{22–23}. In accordance with this, we detected a significant enhancement in soil nutrient cycling due to parasite presence. In particular, we found that the rate of nitrogen (N) mineralization increased markedly ($F_{2,54} = 11.72$, $P < 0.0001$) in the presence of *R. minor* (Fig. 3a); N mineralization increased by 105% and 174% at low and high hemiparasite densities, respectively. As a consequence, the availability of mineral N (DIN) relative to dissolved organic N (DON) also increased, being almost twice as high in the presence of *R. minor* (Fig. 3b). This was due to a significant increase in soil concentrations of mineral N in soil solution, the preferred N source for most plants of temperate grassland communities²⁴. Surprisingly, rates of N mineralization and soil concentrations of DIN did not vary significantly with residual soil fertility, but the concentration of DON in soil solution was found to be lowest ($F_{2,54} = 6.24$, $P < 0.004$) when residual soil fertility was low, perhaps indicating lower N availability in this treatment (data not shown). The lack of difference in DIN in soils of varying residual fertility, however, probably reflects the depletion of available pools of soil inorganic N due to plant growth over the course of the experiment.

The magnitude of the parasite-driven increase in N mineralization

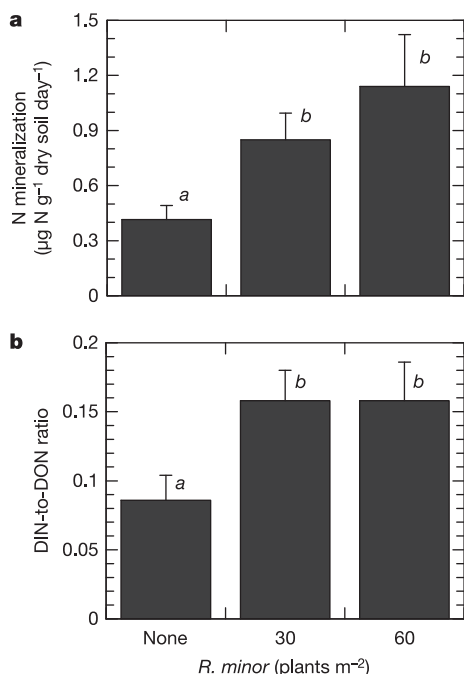


Figure 3 | Mean (±s.e.m.) values for soil N cycling parameters as affected by the presense of *R. minor* after three years. a, Rates of N mineralization. b, Ratio of DIN to DON. Values with different letters are significantly different at the $P < 0.05$ level.

shown here might be expected to feed back to the plant community in terms of increased above-ground productivity and reduced diversity, due to promotion of competitive, fast-growing grasses at the expense of herbs²⁵. That this was not detected points further to the powerful suppressive effect of *R. minor* on competitive, fast-growing grasses and its dominant role as a driver of plant diversity in grassland. In the longer term, however, we expect that indirect effects of parasitic plants on below-ground properties will have far reaching consequences for plant nutrition and production in grassland. The enhancement of soil N cycling in the presence of *R. minor* is most likely an indirect consequence of changes in the quality and quantity of plant inputs to soil, resulting from shifts in the productivity and composition of the sward. It is likely that parasite-driven changes in root growth and turnover, and root exudation patterns, will have altered carbon supply to soil, thereby affecting the activity of decomposer organisms and nutrient mineralization. It has been suggested that parasitic plants might positively influence soil nutrient cycling through their ability to accumulate nutrients in their leaves, thus producing high quality litter that decomposes rapidly, releasing nutrients that would otherwise remain in the host plant or in slowly decomposing litter^{11–13}. This is unlikely to be the case here because the N content of *R. minor* was low compared to other species (see Supplementary Table 4), and the presence of *R. minor* in the sward had no effect on the N content of other species. Also, most above-ground tissue was harvested, so the return of senescent litter to the soil was minimal.

Our findings illustrate that parasitic plants act as major drivers of both the structure and function of grassland ecosystems. They also bolster growing evidence of the importance of above-ground consumers of primary productivity as indirect modulators of ecosystem properties through their influence on below-ground organisms and their activities^{21–23,26}, and specifically illustrate the importance of parasites for understanding ecosystem dynamics⁵. The mechanisms involved in parasite-driven enhancement of nutrient cycling are poorly understood. However, given that almost all terrestrial ecosystems support parasitic plants^{1–4,14}, they are likely to have substantial, but largely unexplored, effects in many ecosystems.

METHODS

Mesocosm design. The mesocosm experiment was set out in March 2000 at the University of Newcastle field station, Close House, England (latitude 54° 59' N, longitude 1° 48' W, elevation 30 m). Mesocosms were 42-l polypropylene pots with a 10-cm base of carboniferous limestone chippings. Each was filled with one of three soils to provide the residual fertility treatments that were a consequence of different past fertilizer practices on the same clayey alluvial soil (Fladbury series) at three sites in Dentdale, Yorkshire (see Supplementary Table 1). Low fertility soil, previously receiving an annual application of farmyard manure, was collected from a traditionally managed hay meadow at Scotchergill Farm (latitude 54° 16' N, longitude 2° 26' W, elevation 138 m). Low-medium fertility soil, previously receiving the same farmyard manure application plus about 25 kg ha⁻¹ of NPK mineral fertilizer, was collected from a hay meadow at Northwaite Farm (latitude 54° 16' N, longitude 2° 25' W, elevation 150 m). A moderate fertility soil, previously receiving slurry and 50 kg ha⁻¹ of NPK mineral fertilizer, was also collected from Northwaite Farm, from a field cut twice a year for silage (latitude 54° 16' N, longitude 2° 24' W, elevation 160 m).

The initial flush of seedlings from the seed bank was removed in May 2000. Mesocosms were then sown with *A. pratensis*, *L. perenne*, *P. bertolonii*, *P. trivialis* and *R. repens*, species characteristic of productive grassland. In September 2000 they were sown with *A. capillaris*, *B. media*, *G. sylvaticum*, *L. corniculatus*, *R. bulbosus* and *R. acris*, species characteristic of traditionally managed northern meadows¹⁸. The *R. minor* treatments (none, 30 and 60 plants m⁻²) were applied in September each year by sowing seeds into appropriate mesocosms. Three farmyard manure treatments (none, 12 tonnes ha⁻¹ alternate years and 12 tonnes ha⁻¹ every year) were based on UK government management prescription for traditional meadows within UK Environmentally Sensitive Areas¹⁹. Composted farmyard manure (total nutrient content: 27.8 mg N g⁻¹; 15.4 mg P g⁻¹; 36.1 mg K g⁻¹) was applied in January of each year. Each treatment combination was replicated three times to give 81 mesocosms, laid out in three blocks. Each of the three treatment replicates was randomly positioned in each block. Artificial autumn grazing was applied each year by randomly

trampling the soil surface of each mesocosm with an artificial 'hoof'. Throughout spring and early summer in every year, all unsown species were removed from the mesocosms, and *R. minor* was thinned out to the required treatment density.

Plant and soil measurements. The sward was cut annually in July and total dry weight of harvested vegetation was recorded. Dry weight of each species was recorded in July 2003. The vegetation within the central 25 × 25 cm grid was cut with shears and hand-sorted to species, expressed as kg ha⁻¹ dry weight. Air-dried plant material of individual species was used for multi-element analysis by a single Kjeldahl micro-digestion²⁴. At the final harvest, roots of selected plant species were taken for determination of root colonization by arbuscular mycorrhizal fungi²⁷. In July 2003, five soil samples were taken from each mesocosm, bulked, passed through a 4-mm sieve, and stored at 4°C until analysis. Soil microbial community structure of these samples was assessed using phospholipid fatty acid analysis²⁸, and soil N availability was measured using three techniques²⁴. First, concentrations of mineral N (DIN) in soil solution were determined by auto-analysis. Second, net N mineralization was measured as the release of mineral N (NH₄⁺-N and NO₃⁻-N) after incubation of samples for 14 days at 25°C. After incubation, soil concentrations of mineral N were determined. Third, dissolved organic N (DON) was measured by oxidation of water extracts with potassium persulphate (K₂S₂O₈), and measurement of the resultant mineral N by autoanalyser procedures.

Data analysis. Main treatment effects on biomass of the most frequent plant species, and on the total biomass each year, were assessed by analysis of variance (ANOVA) using MINITAB. Interactions between treatments were tested as part of the ANOVA. Where a significant difference was found with the ANOVA test, the significance of differences between means was tested using Tukey's test. When a species was absent from one treatment, analysis was based on a reduced number of treatments. Analysis of variance was used to assess treatment differences for soil biological variables. This was also done for the N, P and K concentration in the above-ground parts of the main plant species. A test of normality, the Anderson–Darling test, was applied to the residuals of each analysis. If this deviated significantly from normality then a suitable transformation was applied to the original data. The proportion of the variability in 2003 species composition attributable to each treatment, and its significance, were based on ordination methods using CANOCO²⁹. The same data for yield and diversity were based on the sums of squares associated with each treatment in an ANOVA test.

Received 17 July; accepted 5 September 2005.

- Kuijt, J. *The Biology of Parasitic Flowering Plants* (Univ. California Press, Berkeley, 1969).
- Press, M. C. & Graves, J. D. *Parasitic Plants* (Chapman and Hall, London, 1995).
- Press, M. C. Dracula or Robin Hood? A functional role for root hemiparasites in nutrient poor ecosystems. *Oikos* **82**, 609–611 (1998).
- Parker, C. A. & Riches, C. R. *Parasitic Weeds of the World: Biology and Control* (CAB International, Wallingford, 1993).
- Thomas, F., Renaud, F. & Guégan, J. F. *Parasites in the Grand Scheme of Things* (Oxford Univ. Press, Oxford, 2005).
- Gibson, C. C. & Watkinson, A. R. The role of the hemiparasitic annual *Rhinanthus minor* in determining grassland community structure. *Oecologia* **89**, 62–68 (1992).
- Pennings, S. C. & Callaway, R. M. Impact of a parasitic plant on the structure and dynamics of salt marsh vegetation. *Ecology* **77**, 1410–1419 (1996).
- Davies, D. M., Graves, J. D., Elias, C. O. & Williams, P. J. The impact of *Rhinanthus* spp. on sward productivity and composition: implications for the restoration of species-rich grassland. *Biol. Conserv.* **82**, 87–93 (1997).
- Marvier, M. A. Parasite impacts on host communities: plant parasitism in a Californian coastal prairie. *Ecology* **79**, 2616–2623 (1998).
- Smith, D. The population dynamics and community ecology of root hemiparasitic plants. *Am. Nat.* **155**, 13–23 (2000).
- Questaed, H. M. *et al.* Decomposition of sub-arctic plants with differing nitrogen economies: a functional role for hemiparasites. *Ecology* **84**, 3209–3221 (2003).
- Questaed, H. M., Press, M. C. & Callaghan, T. V. Litter of the hemiparasite *Bartsia alpina* enhances plant growth: evidence for a functional role in nutrient cycling. *Oecologia* **135**, 606–614 (2003).
- Questaed, H. M., Press, M. C., Callaghan, T. V. & Cornelissen, J. H. C. The hemiparasite angiosperm *Bartsia alpina* has the potential to accelerate decomposition in sub-arctic communities. *Oecologia* **130**, 88–95 (2002).
- Westbury, D. B. *Rhinanthus minor* L. *J. Ecol.* **63**, 335–344 (2004).
- Grime, J. P., Hodgson, J. G. & Hunt, R. *The Abridged Comparative Plant Ecology* (Unwin Hyman, London, 1990).
- Gibson, C. C. & Watkinson, A. R. Host selectivity and the mediation of competition by the root hemiparasite *Rhinanthus minor*. *Oecologia* **86**, 81–87 (1991).
- Pywell, R. F. *et al.* Facilitating grassland diversification using the hemiparasitic plant *Rhinanthus minor*. *J. Appl. Ecol.* **41**, 880–887 (2004).
- Rodwell, J. S. *Grasslands and Montane Communities* (Cambridge Univ. Press, Cambridge, 1992).
- Smith, R. S. *et al.* Soil microbial community, fertility, vegetation and diversity as targets in the restoration of a meadow grassland. *J. Appl. Ecol.* **40**, 51–64 (2003).
- Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* 2nd edn (Academic, London, 1997).
- Bardgett, R. D. & Wardle, D. A. Herbivore mediated linkages between aboveground and belowground communities. *Ecology* **84**, 2258–2268 (2002).
- Wardle, D. A. *et al.* Ecological linkages between aboveground and belowground biota. *Science* **304**, 1629–1633 (2004).
- Bardgett, R. D. *The Biology of Soil: A Community and Ecosystem* (Oxford Univ. Press, Oxford, 2005).
- Bardgett, R. D., Streeter, T. & Bol, R. Soil microbes compete effectively with plants for organic nitrogen inputs to temperate grasslands. *Ecology* **84**, 1277–1287 (2003).
- Stevens, C. J., Dise, N. B., Mountford, J. O. & Gowing, D. J. Impact of nitrogen deposition on the species richness of grasslands. *Science* **303**, 1876–1879 (2004).
- McNaughton, S. J., Banyikwa, F. F. & McNaughton, M. M. Promotion of the cycling of diet-enhancing nutrients by African grazers. *Science* **278**, 1798–1800 (1997).
- McGonigle, T. P., Miller, M. H., Evans, D. G., Fairchild, G. L. & Swan, J. A. A new method which gives an objective measure of colonisation of roots by vesicular-arbuscular mycorrhizal fungi. *New Phytol.* **115**, 495–501 (1990).
- Bardgett, R. D., Hobbs, P. J. & Frostegård, Å. Changes in fungal:bacterial biomass ratios following reductions in the intensity of management on an upland grassland. *Biol. Fertil. Soils* **22**, 261–264 (1996).
- ter Braak, C. J. F. & Smilauer, P. *CANOCO Reference Manual and CANODRAW for Windows User's Guide: Software for Canonical Community Ordination (version 4.5)* 265–268 (Microcomputer Power, Ithaca, New York, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the UK Department for Environment, Food and Rural Affairs (DEFRA) for funding this work under their species-rich grassland programme. We thank D. Wardle, L. Walker, E. Ayres, R. Van der Wal, K. Harrison, N. Ostle and K. Wilson for their comments on the manuscript, and R. Brand Hardy of DEFRA for his continued support of this work.

Author Contributions R. S. Smith was the project leader, responsible for the experimental design and the vegetation aspects, with S.P. and J.M.S. responsible for most of the data collection and management of the mesocosm experiment. R.D.B. was responsible for below-ground aspects of the study, with assistance from H.Q. and P.J.H. R. S. Shiel provided statistical support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D.B. (r.bardgett@lancaster.ac.uk).

Facioscapulohumeral muscular dystrophy in mice overexpressing *FRG1*

Davide Gabellini¹, Giuseppe D'Antona², Maurizio Moggio³, Alessandro Prella³, Chiara Zecca³, Raffaella Adami², Barbara Angeletti^{4,†}, Patrizia Ciscato³, Maria Antonietta Pellegrino², Roberto Bottinelli², Michael R. Green¹ & Rossella Tupler^{1,4,†}

Facioscapulohumeral muscular dystrophy (FSHD) is an autosomal dominant neuromuscular disorder that is not due to a classical mutation within a protein-coding gene^{1,2}. Instead, almost all FSHD patients carry deletions of an integral number of tandem 3.3-kilobase repeat units, termed D4Z4, located on chromosome 4q35 (ref. 3). D4Z4 contains a transcriptional silencer whose deletion leads to inappropriate overexpression in FSHD skeletal muscle of 4q35 genes located upstream of D4Z4 (ref. 4). To identify the gene responsible for FSHD pathogenesis, we generated transgenic mice selectively overexpressing in skeletal muscle the 4q35 genes *FRG1*, *FRG2* or *ANT1*. We find that *FRG1* transgenic mice develop a muscular dystrophy with features characteristic of the human disease; by contrast, *FRG2* and *ANT1* transgenic mice seem normal. *FRG1* is a nuclear protein and several lines of evidence suggest it is involved in pre-messenger RNA splicing^{5–7}. We find that in muscle of *FRG1* transgenic mice and FSHD patients, specific pre-mRNAs undergo aberrant alternative splicing. Collectively, our results suggest that FSHD results from inappropriate overexpression of *FRG1* in skeletal muscle, which leads to abnormal alternative splicing of specific pre-mRNAs.

We generated transgenic mice overexpressing *FRG1*, *FRG2* (for FSHD region genes 1 and 2) or *ANT1* (for adenine nucleotide translocase type 1; also known as SLC25A4) under the control of the human skeletal α -actin (HSA) gene promoter⁸. Genotypes were verified by polymerase chain reaction (PCR) and Southern blot analysis (data not shown), and PCR with reverse transcription (RT-PCR) confirmed that the transgenes were expressed selectively in skeletal muscle (Supplementary Fig. S1a). Mice overexpressing *FRG1* had a strikingly altered phenotype (see below). By contrast, mice overexpressing *FRG2* or *ANT1* lacked obvious phenotypic abnormalities (Supplementary Fig. S1b).

The number of D4Z4 repeats is a critical determinant of FSHD disease severity: fewer repeats are associated with a more severe phenotype⁹. We have previously shown that the extent of 4q35 gene overexpression in FSHD skeletal muscle is inversely related to the number of D4Z4 repeats, suggesting that 4q35 gene overexpression directly correlates with disease severity⁴. To determine whether this was also the case in *FRG1* mice, we selected three lines in which the transgene was overexpressed (relative to endogenous mouse *Frg1*) at levels within the range previously reported for *FRG1* overexpression in FSHD muscle biopsies⁴; we refer to these lines as *FRG1*-low, *FRG1*-med, and *FRG1*-high (Fig. 1a–c). *FRG1* expression levels were also quantified by real-time PCR and transgene copy number was determined by Southern blot analysis (Supplementary Fig. S2). Figure 1d shows that the degree of severity of kyphosis (an abnormal

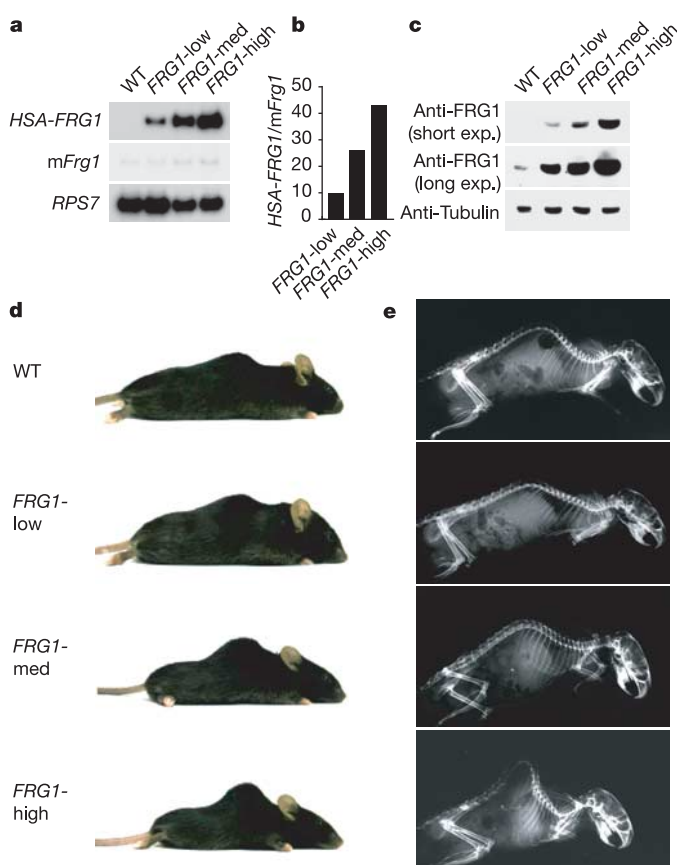


Figure 1 | Mice overexpressing *FRG1* show kyphosis. **a**, RT-PCR analysis monitoring expression of the *FRG1* transgene (*HSA-FRG1*), the endogenous mouse *Frg1* gene (*mFrg1*), and the ribosomal gene *RPS7* (as a loading control) in vastus muscle from wild-type (WT) mice and three transgenic lines overexpressing *FRG1* at different levels: *FRG1*-low, *FRG1*-med and *FRG1*-high. **b**, Quantification of *FRG1* transgene expression relative to the endogenous *mFrg1* gene of samples shown in **a**. **c**, Immunoblot analysis of *FRG1* protein expression in vastus muscle from WT and *FRG1* transgenic mice. Tubulin was monitored as a loading control. **d**, Photomicrographs of WT and *FRG1* transgenic mice showing kyphosis. **e**, Radiograms of the skeletons of *FRG1* transgenic mice.

¹Howard Hughes Medical Institute, Programs in Gene Function and Expression and Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. ²Dipartimento di Medicina Sperimentale, Sezione di Fisiologia umana, Università degli Studi di Pavia, 27100 Pavia, Italy. ³U.O. Neurologia, Università degli Studi di Milano, I.R.C.C.S. Fondazione Ospedale Maggiore, 20122 Milano, Italy. ⁴Biologia Generale e Genetica Medica, Università degli Studi di Pavia, 27100 Pavia, Italy. [†]Present address: Dipartimento di Scienze Biomediche, Università degli Studi di Modena e Reggio Emilia, 41100 Modena, Italy.

outward curvature of the spine) correlated with transgene expression levels: *FRG1*-low mice showed no evidence of kyphosis, whereas *FRG1*-med and *FRG1*-high mice exhibited mild and severe kyphosis, respectively. Radiograms of the skeletons of *FRG1* transgenic mice showed no evidence of skeletal deformities (Fig. 1e), suggesting that the spinal abnormality was due to muscle weakness. Notably, abnormal spinal curvature due to muscle weakness is found in FSHD patients^{1,10} as well as mouse models of other muscular dystrophies^{11–13}.

Skeletal muscle from *FRG1* mice showed dystrophic histological and ultrastructural features characterized by increased fibre size variability, fibre necrosis, number of central nuclei, and connective tissue (Fig. 2 and summarized in Supplementary Table 1). Quantification of the number of centrally nucleated fibres showed a moderate level of central nuclei in *FRG1* transgenic mice relative to *mdx* mice, a model for Duchenne muscular dystrophy (DMD) (Supplementary Fig. S3). The extent of dystrophy correlated with *FRG1* overexpression levels (compare Fig. 2e–g with i–k and m–o). Notably, different muscle types were affected to various extents and showed variable histopathological changes (compare Fig. 2e–m with f–n and g–o; quantified in Supplementary Table 1), similar to what has been observed in FSHD patients². Moreover, for all muscle types, histopathological signs of muscular dystrophy increased with age (data not shown), again consistent with the human disease. Electron microscopy revealed a marked proliferation of collagen fibrils in *FRG1*-med and *FRG1*-high mice compared with *FRG1*-low mice (compare Fig. 2l, p with h). Other signs of muscular pathology were also evident in *FRG1* mice; however, in all muscles examined there was no alteration of mitochondrial activity (Supplementary Fig. S4 and Supplementary Table 1). As expected, abnormal muscle histology was not detected in mice overexpressing *FRG2* or *ANT1* (Supplementary Fig. S5). Collectively, these results indicate that mice overexpressing *FRG1* develop a progressive muscular dystrophy whose degree of severity correlates with the level of transgene overexpression.

FRG1 overexpression also correlated with a reduction in body

weight. Figure 3a shows that *FRG1*-low mice were indistinguishable in weight compared to wild-type littermates, whereas *FRG1*-med and *FRG1*-high mice were considerably smaller (~80% and ~60% of wild type, respectively). Food consumption relative to body weight was not significantly different between wild-type and *FRG1* mice, suggesting that reduced body weight was not due to decreased caloric intake (Fig. 3b).

To determine whether the reduction in body weight was due to muscle atrophy, we analysed muscle weight in *FRG1* mice. A reduction in muscle mass in *FRG1*-med and *FRG1*-high mice was obvious on visual observation (Fig. 3c). Indeed, total muscle weight was significantly lower in *FRG1*-med (~75% of wild type) and *FRG1*-high (~50% of wild type) mice than in *FRG1*-low mice (Fig. 3d). Measurements of single muscle fibre cross-sectional area (CSA), an assessment of fibre size, indicated that a principal determinant of muscle weight reduction in *FRG1*-med and *FRG1*-high mice was fibre atrophy (Fig. 3e). The extent of the reduction in muscle weight in *FRG1*-med and *FRG1*-high mice varied widely among different muscle types (Supplementary Fig. S6a) and correlated with a decrease in CSA (Supplementary Fig. S6b). Notably, the weight of other organs did not significantly differ between wild-type and *FRG1* mice (Supplementary Fig. S6c). Collectively, these results indicate that the reduced body weight of *FRG1* mice is primarily due to selective muscle atrophy.

To determine whether the muscle atrophy observed in *FRG1* mice affected muscle function, we performed exercise tolerance tests. Animals underwent a weekly treadmill test for an 18-week period and were monitored for the time taken to reach exhaustion. As expected, over the entire monitoring period, *FRG1*-med and *FRG1*-high mice showed, on average, reduced tolerance to exercise compared with wild-type mice, whereas *FRG1*-low mice were indistinguishable from wild types (Fig. 3f). Notably, *FRG1*-high mice showed a comparable reduction in exercise tolerance to *mdx* mice.

Most muscular dystrophies have been linked to mutations in genes encoding proteins of the dystrophin–glycoprotein complex (DGC)¹⁴, or for enzymes probably involved in the post-translational

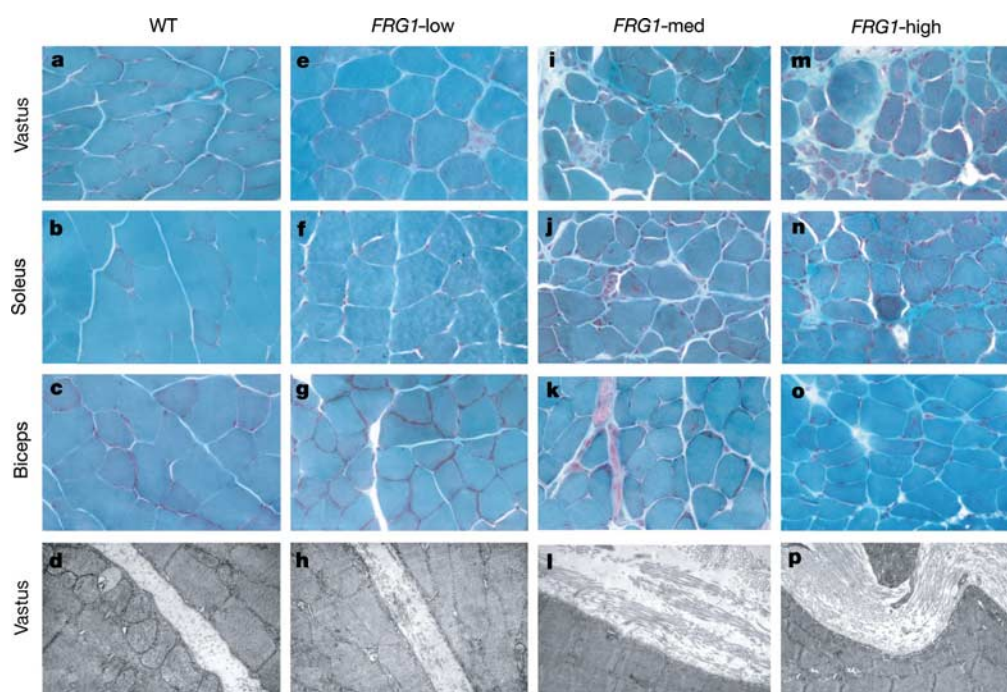


Figure 2 | Skeletal muscles from *FRG1* transgenic mice show dystrophic histological features. Gomori trichrome-stained cross-cryosections (original magnification, $\times 400$) of vastus, soleus and biceps muscle from WT (a–c), *FRG1*-low (e–g), *FRG1*-med (i–k) and *FRG1*-high (m–o) transgenic

mice at 13 weeks of age. Electron micrographs (original magnification, $\times 7,500$) of vastus muscle from WT (d), *FRG1*-low (h), *FRG1*-med (l) and *FRG1*-high (p) mice stained with uranyl acetate and lead citrate are also shown.

modification of these proteins¹⁵. Accordingly, alteration of sarcolemmal integrity has been described in nearly all mouse models of muscular dystrophy^{14,16}. We analysed sarcolemmal integrity in *FRG1* mice by three independent methods: Evans blue dye staining, serum levels of creatine kinase, and immunofluorescence analysis of DGC components. Consistent with the human FSHD phenotype, *FRG1* transgenic mice showed no alteration of sarcolemmal integrity (Supplementary Fig. S7).

Several lines of evidence suggest that *FRG1* plays a role in RNA processing⁵⁻⁷. For example, *FRG1* is a nuclear protein that assumes an intranuclear 'speckled' pattern (ref. 7 and Supplementary Fig. S8a) characteristic of a number of mammalian splicing factors, and proteomic studies have found that *FRG1* is a component of purified

spliceosomes⁶. These observations raised the possibility that alterations in *FRG1* levels could affect pre-mRNA splicing. In this regard, abnormal RNA processing is associated with several neuromuscular disorders¹⁷⁻¹⁹, and, in particular, in myotonic dystrophy (DM) alternative splicing of specific pre-mRNAs is misregulated¹⁹. To test whether *FRG1* overexpression leads to aberrant alternative splicing, we analysed the alternative splicing patterns of three genes that are aberrantly spliced in myotonic dystrophy patients and animal models: fast skeletal muscle troponin T (*Tnnt3*), myotubularin related protein 1 (*Mtmr1*) and skeletal muscle chloride channel 1 (*Clcn1*)^{20,21}. Figure 4a-c shows that in skeletal muscle of *FRG1* transgenic mice, the alternative splicing pattern of *Tnnt3* and

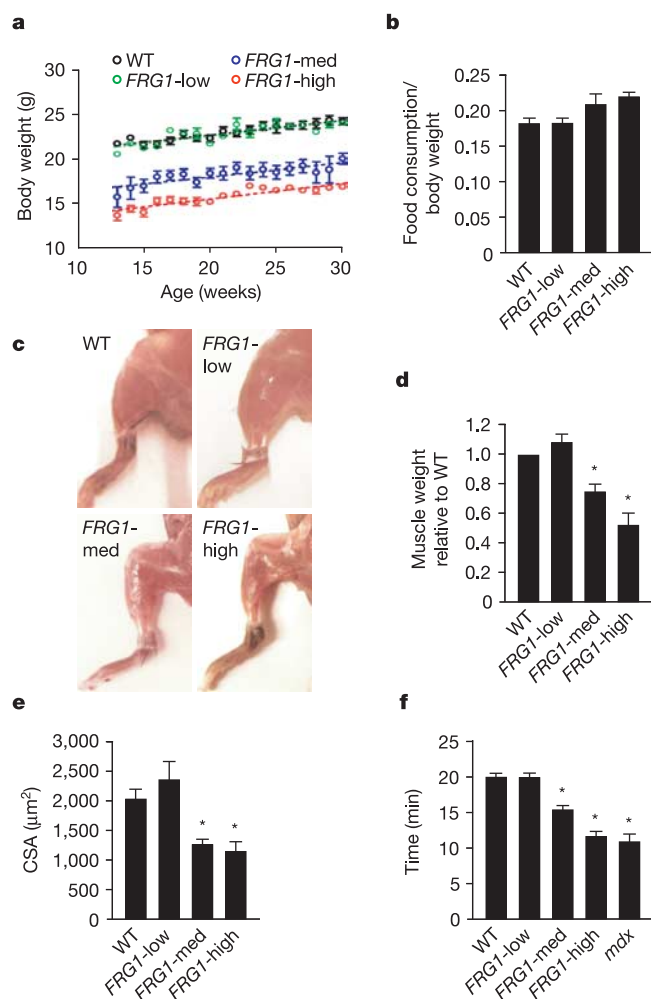


Figure 3 | Increased *FRG1* transgene expression correlates with a reduction in body weight and muscle atrophy. **a**, Time course of body weight gain in wild-type (WT) mice and *FRG1* transgenic mice ($n = 6$ mice per group). Animals were monitored weekly from 13–30 weeks of age. **b**, Ratio of food consumption to body weight of WT and *FRG1* transgenic mice ($n = 6$ mice per group) at 13 weeks of age. Values are means $\pm$ s.e.m.; asterisks indicate significant differences ($P < 0.05$) from WT. **c**, Photomicrographs of muscle from WT and *FRG1* transgenic mice at 13 weeks of age. **d**, Relative muscle weight of WT and *FRG1* transgenic mice. Weights of seven muscle groups ($n = 8$ for soleus, tibialis anterior, vastus lateralis, gastrocnemius, triceps and biceps; $n = 4$ for diaphragm) from *FRG1* transgenic mice were averaged and expressed relative to the weight of the muscles in WT animals. **e**, Mean cross-sectional area (CSA) of individual fibres from vastus, soleus, diaphragm and biceps muscles ($n = 3$ per muscle type; $n = 3$ mice per group). **f**, Time taken to reach exhaustion following treadmill exercise for WT and *FRG1* transgenic mice ($n = 6$ per group) and *mdx* mice ($n = 3$); the average value for weeks 15–30 is shown.

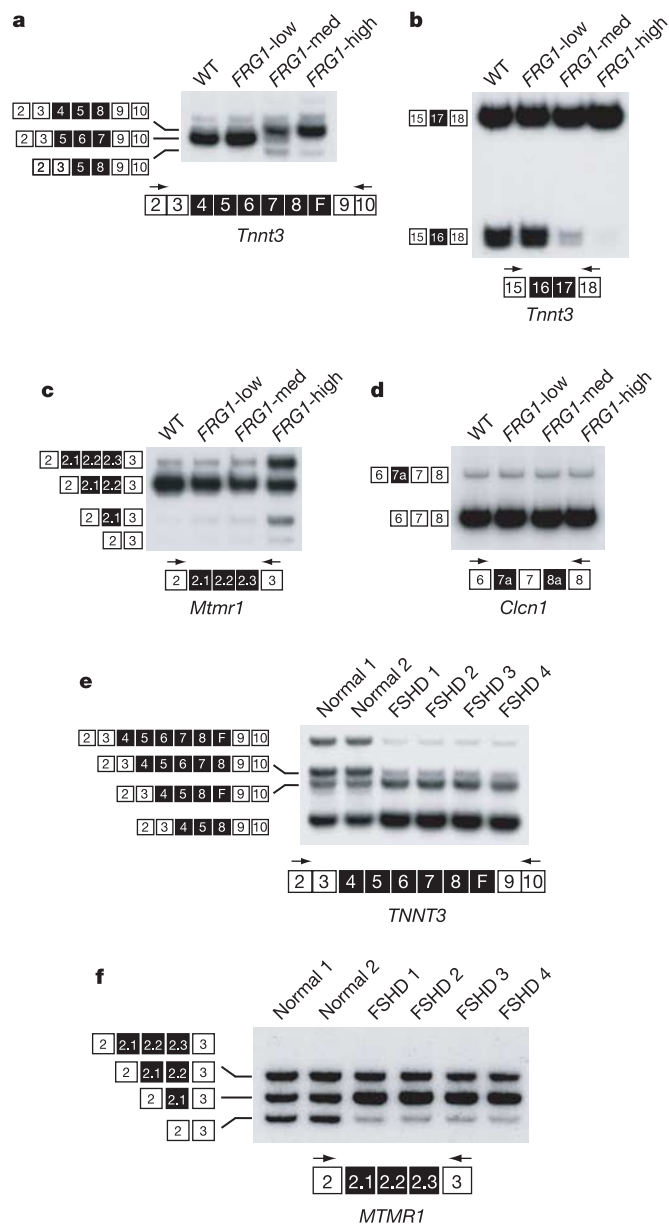


Figure 4 | Muscles from *FRG1* transgenic mice and FSHD patients show aberrant pre-mRNA splicing of selective genes. **a–d**, RT-PCR analysis of *Tnnt3* (**a**, **b**), *Mtmr1* (**c**) and *Clcn1* (**d**) splicing in vastus muscle from 13-week-old WT or *FRG1* transgenic mice. Open boxes illustrate constitutive exons, black boxes indicate alternatively spliced exons, and arrows denote primer positions. Identities of alternatively spliced products are indicated to the left of the gels. For *Tnnt3*, 'F' denotes the fetal exon. **e**, **f**, Analysis of *TNNT3* (**e**) and *MTMR1* (**f**) splicing in muscle cell cultures derived from two normal individuals (Normal 1, 2) and four FSHD patients (FSHD 1–4).

Mtmr1 differed from that of wild-type mice. Furthermore, the extent of aberrant splicing correlated with the level of *FRG1* overexpression. By contrast, splicing of *Cln1* was unaffected by *FRG1* overexpression (Fig. 4d). Significantly, aberrant splicing of *Tnnt3* and *Mtmr1* was not detected in mice overexpressing *FRG2* or *ANT1* (Supplementary Fig. S8b–d).

Several observations indicate that the aberrant splicing observed in *FRG1* mice was not a secondary effect of the dystrophic process. First, aberrant splicing of *Tnnt3* and *Mtmr1* could be detected in 4-week-old *FRG1*-high mice (Supplementary Fig. S8e–g), a time at which histological and ultrastructural analyses revealed no evidence for muscular dystrophy. Second, *Tnnt3* and *Mtmr1* were spliced normally in *mdx* mice (Supplementary Fig. S8e–g). Finally, altered *Tnnt3* and *Mtmr1* splicing was also observed in C2C12 muscle cell cultures overexpressing *FRG1* (Supplementary Fig. S8h–i) but not *ASF/SF2* (alternate splicing factor/splicing factor 2; Supplementary Fig. S8j–k), indicating that aberrant splicing was a direct and specific effect of *FRG1* overexpression.

The results in *FRG1* mice prompted us to analyse splicing in FSHD patients. Figure 4e, f shows that splicing of *TNNT3* and *MTMR1* was altered in muscle cell cultures derived from patients with FSHD, in which *FRG1* was overexpressed (Supplementary Fig. S9a). By contrast, splicing of *TNNT3* and *MTMR1* was normal in muscle cell cultures derived from patients with DMD and congenital merosin-deficient muscular dystrophy 1A (Supplementary Fig. S9b), in which *FRG1* expression was normal (Supplementary Fig. S9c). We note that the alternative splicing patterns in FSHD muscle cell cultures differed from those observed in *FRG1* mice, which most likely reflects differences in splicing between humans and mice, and between intact muscle and cultured muscle cells^{20,21}.

In this study, we have shown that transgenic overexpression of *FRG1* in skeletal muscle causes muscular dystrophy in mice. To our knowledge, this is the first report of a genetically inherited disease caused by inappropriate overexpression of a single wild-type gene. By contrast, mice overexpressing two other putative FSHD candidate genes, *FRG2* and *ANT1*, are normal with regard to both phenotype and muscle histology. Notably, *FRG1* mice show a variety of features found in FSHD, including abnormal spinal curvature, progressive muscle dystrophy, and skeletal muscle atrophy. In addition, *FRG1* mice show a differential involvement of muscle types that is strikingly similar to that observed in FSHD but not in other muscular dystrophies^{1,2}. In fact, there is almost a complete correlation in the types of muscles affected and the degree to which they are affected, with the exception of the gastrocnemius (Supplementary Table 2). Finally, in both FSHD patients and *FRG1* mice there is no evidence for mitochondrial involvement²² or alteration of sarcolemmal integrity²³. This latter feature distinguishes FSHD from other muscular dystrophies in which sarcolemmal disruption is the primary pathogenetic mechanism¹⁴. Our proposal that FSHD is caused by *FRG1* overexpression is supported by studies of individuals harbouring large genomic deletions in the 4q35 region. For example, some FSHD patients carry deletions that include *FRG2* and a portion of *D4Z4* (ref. 24), indicating that *FRG2* is dispensable for disease pathogenesis. Notably, individuals carrying deletions of *D4Z4*, *FRG2* and *FRG1* have no phenotypic consequences²⁵, consistent with a requirement of *FRG1* for FSHD.

We found changes in the alternative splicing pattern of two pre-mRNAs, *TNNT3* and *MTMR1*, in muscle of *FRG1* mice and FSHD patients. Altered splicing of these pre-mRNAs may contribute to FSHD: *TNNT3*, which regulates muscle contractility²⁶, and *MTMR1*, which can regulate muscle atrophy²⁰, are also aberrantly spliced in myotonic dystrophy and have been proposed to contribute to disease phenotype¹⁹. However, more than 75% of human genes undergo alternative splicing²⁷, and thus in muscle cells overexpressing *FRG1*, splicing of hundreds-to-thousands of pre-mRNAs may be altered. It is likely that the cumulative effect of decreased levels of many normal protein isoforms and increased levels of many aberrant protein

isoforms is responsible for disease. The development of a faithful animal model of FSHD should provide further insights into the molecular basis and pathogenesis of the disease, as well as a model for evaluating therapeutic strategies.

METHODS

Generation of transgenic mice. *FRG1*, *FRG2* and *ANT1* open reading frames were PCR-amplified using the following primers: *FRG1*, HSA-*FRG1*-F 5'-GATCTAGCGGCCGCCATGGCCGAGTACTCCTATGTGAAGTCT-3' and HSA-*FRG1*-R 5'-GCGCGCTTAATTAATCACTTGCAGTATCTGTCGGCTTCA-3'; *FRG2*, HSA-*FRG2*-F 5'-GATCTAGCGGCCGCCAATGGGAAAGGGAAATGAAGACTCCGA-3' and HSA-*FRG2*-R 5'-GCGCGCTTAATTAATCAATCCAGAGCTGCATCTCTGCT-3'; *ANT1*, HSA-*ANT1*-F 5'-GATCTAGCGGCCGCCATGGGTGATCAGCTTGGAGCTT-3' and HSA-*ANT1*-R 5'-GCGCGCTTAATTAATAGACATATTTTGTATCTCATCAT-3'. PCR products were sequenced and cloned into pBSX-HSAvpA, provided by J. Chamberlain. Transgenes were excised from agarose gels and purified for microinjection into fertilized eggs recovered from C57BL/6 females crossed with C57BL/6 males. All procedures were performed at the University of Massachusetts Medical School Transgenic Animal Modelling Core Facility.

Analysis of *FRG1* transgene expression. Frozen tissue was pulverized using a mortar and pestle (Coors) under liquid nitrogen. Radioactive RT-PCR was performed as previously described⁴ using 1 µg of purified DNA-free RNA and the following primers: HSA-*FRG1*, HSA-RT 5'-CGAGCCGAGAGTAGCAGTTG-3' and HSA-*FRG1*-R 5'-GCGCGCTTAATTAATCACTTGCAGTATCTGTCGGCTTCA-3'; HSA-*FRG2*, HSA-RT 5'-CGAGCCGAGAGTAGCAGTTG-3' and HSA-*FRG2*-R 5'-GCGCGCTTAATTAATCACTTCCAGAGCTGCATCTCTGCT-3'; HSA-*ANT1*, HSA-RT 5'-CGAGCCGAGAGTAGCAGTTG-3' and HSA-*ANT1*-R 5'-GCGCGCTTAATTAATAGACATATTTTGTATCTCATCAT-3'; mouse ribosomal protein S7, mRPS7-F 5'-CAGCTGCGTGAACCTCAACAT-3' and mRPS7-R 5'-TCACACGGATCCTCTTACCC-3'. For protein analysis, proteins were extracted using 3 ml of ice-cold RIPA buffer per gram of pulverized tissue. Samples were incubated for 30 min on ice, briefly sonicated, and centrifuged at 16,500 g for 10 min at 4 °C. The supernatant was removed and centrifuged again. Protein concentration of the final supernatant was measured using the BCA assay (Pierce). Immunoblot analysis was performed as previously described⁴ using 30 µg of extract and affinity-purified rabbit anti-*FRG1* antibodies (generated by AnaSpec using the peptide KDKKRKREDEETQ) and a monoclonal anti-tubulin antibody (Sigma).

Morphological analysis. Skeletal muscle tissue specimens were frozen in isopentane cooled in liquid nitrogen, and stored in liquid nitrogen until use. Cross-cryosections (8-µm thick) were prepared and stained with Gomori trichrome as previously described²⁸. Ultrastructural studies were performed as previously described²⁸.

Body weight, food consumption and muscle weight analysis. Mice (C57BL/6 females) were housed individually and given unlimited access to food and water for the duration of the study. Four groups of mice ($n = 6$ per group)—wild type, *FRG1*-low, *FRG1*-med and *FRG1*-high—were monitored weekly from 13–30 weeks of age for body weight, and food and water consumption. For muscle weight analysis, the diaphragm, soleus, tibialis, vastus, gastrocnemius, triceps and biceps muscles from 13-week-old mice were dissected using a stereomicroscope, blotted on filter paper, and weighed.

Analysis of single fibre cross sectional area (CSA). The CSA of single muscle fibres from vastus, soleus, diaphragm and biceps muscles ($n = 3$ per muscle type; $n = 3$ mice per group) was determined by Scion Image software (NIH) on calibrated images of cross-cryosections stained with haematoxylin and eosin. CSA was determined at $\times 320$ magnification.

Exercise tolerance tests. Four groups of mice ($n = 6$ per group)—wild type, *FRG1*-low, *FRG1*-med and *FRG1*-high—were subjected weekly to an exhaustion treadmill test from 13–30 weeks of age. Each mouse was placed on the belt of a six-lane motorized treadmill (Exer 3/6 Treadmill, Columbus Instruments) supplied with shocker plates. The treadmill was run at an inclination of 0° at 5 m min⁻¹ for 5 min, after which time the speed was increased 1 m min⁻¹ every minute. The test was stopped when the mouse remained on the shocker plate for more than 20 s without attempting to re-engage the treadmill, and the time to exhaustion was determined.

Cell lines and cell culture. The C2C12 mouse myoblast cell line (ATCC) was cultured in standard conditions. To generate myotubes, 80% confluent cells were shifted to differentiation medium (DMEM supplemented with 2% horse serum, 2 mM glutamine and 70 µg ml⁻¹ gentamycin) and were maintained for six days with the medium changed every day.

Overexpression of *FRG1* in C2C12 cells. *FRG1* was cloned into the pFLAG-HA-IRES-Neo (pFH) vector, provided by R. G. Roeder. C2C12 cells were

transfected with linearized pFH-FRG1 or pFH using Effectene (Qiagen) according to the manufacturer's instructions, and 48 h later the cells were passed and G418 (1 mg ml^{-1} , Calbiochem) was added to the media. G418-resistant cells were maintained as a pool and grown under constant selection. FRG1 expression was confirmed by immunoblotting and immunofluorescence.

Analysis of mRNA alternative splicing. Total RNA was purified from frozen tissue as described above, or from cell lines using RNeasy (Qiagen). For C2C12 studies, cells were grown to confluence and then shifted to differentiation media (DMEM, 2% horse serum, 1 mg ml^{-1} G418) and maintained for six days with daily media changes. After six days, total RNA was purified using RNeasy and analysed by RT-PCR. RT-PCR primers and conditions were previously described^{20,21}. PCR products were gel purified and sequenced.

Received 10 August; accepted 11 November 2005.

Published online 11 December 2005.

- Tawil, R. Facioscapulohumeral muscular dystrophy. *Curr. Neurol. Neurosci. Rep.* **4**, 51–54 (2004).
- Flanigan, K. M. in *Myology* (eds Engel, A. & Franzini-Armstrong, C.) 1123–1133 (McGraw Hill Professional, New York, 2004).
- van Deutekom, J. C. et al. FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* **2**, 2037–2042 (1993).
- Gabellini, D., Green, M. R. & Tupler, R. Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell* **110**, 339–348 (2002).
- Kim, S. K. et al. A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
- Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231–1245 (2002).
- van Koningsbruggen, S. et al. FRG1P is localised in the nucleolus, Cajal bodies, and speckles. *J. Med. Genet.* **41**, e46 (2004).
- Muscat, G. E. & Kedes, L. Multiple 5'-flanking regions of the human α -skeletal actin gene synergistically modulate muscle-specific expression. *Mol. Cell. Biol.* **7**, 4089–4099 (1987).
- Ricci, E. et al. Progress in the molecular diagnosis of facioscapulohumeral muscular dystrophy and correlation between the number of *KpnI* repeats at the 4q35 locus and clinical phenotype. *Ann. Neurol.* **45**, 751–757 (1999).
- Kilmer, D. D. et al. Profiles of neuromuscular diseases. Facioscapulohumeral muscular dystrophy. *Am. J. Phys. Med. Rehabil.* **74**, S131–S139 (1995).
- Deconinck, A. E. et al. Utrophin-dystrophin-deficient mice as a model for Duchenne muscular dystrophy. *Cell* **90**, 717–727 (1997).
- Grady, R. M. et al. Skeletal and cardiac myopathies in mice lacking utrophin and dystrophin: a model for Duchenne muscular dystrophy. *Cell* **90**, 729–738 (1997).
- Lynch, G. S., Hinkle, R. T., Chamberlain, J. S., Brooks, S. V. & Faulkner, J. A. Force and power output of fast and slow skeletal muscles from *mdx* mice 6–28 months old. *J. Physiol. (Lond.)* **535**, 591–600 (2001).
- Durbeej, M. & Campbell, K. P. Muscular dystrophies involving the dystrophin-glycoprotein complex: an overview of current mouse models. *Curr. Opin. Genet. Dev.* **12**, 349–361 (2002).
- Muntoni, F., Brockington, M. & Brown, S. C. Glycosylation eases muscular dystrophy. *Nature Med.* **10**, 676–677 (2004).
- Dalkilic, I. & Kunkel, L. M. Muscular dystrophies: genes to pathogenesis. *Curr. Opin. Genet. Dev.* **13**, 231–238 (2003).
- Calado, A. et al. Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. *Hum. Mol. Genet.* **9**, 2321–2328 (2000).
- Gubitz, A. K., Feng, W. & Dreyfuss, G. The SMN complex. *Exp. Cell Res.* **296**, 51–56 (2004).
- Ranum, L. P. & Day, J. W. Myotonic dystrophy: RNA pathogenesis comes into focus. *Am. J. Hum. Genet.* **74**, 793–804 (2004).
- Buj-Bello, A. et al. Muscle-specific alternative splicing of myotubularin-related 1 gene is impaired in DM1 muscle cells. *Hum. Mol. Genet.* **11**, 2297–2307 (2002).
- Kanadia, R. N. et al. A muscleblind knockout model for myotonic dystrophy. *Science* **302**, 1978–1980 (2003).
- Orrell, R. W. et al. Definitive molecular diagnosis of facioscapulohumeral dystrophy. *Neurology* **52**, 1822–1826 (1999).
- Rogers, M., Sewry, C. A. & Upadhyaya, M. in *FSHD Facioscapulohumeral Muscular Dystrophy: Clinical Medicine and Molecular Cell Biology* (eds Upadhyaya, M. & Cooper, D. N.) 277–298 (BIOS Scientific Publishers, London and New York, 2004).
- Lemmers, R. J. et al. D4F104S1 deletion in facioscapulohumeral muscular dystrophy: phenotype, size, and detection. *Neurology* **61**, 178–183 (2003).
- Tupler, R. et al. Monosomy of distal 4q does not cause facioscapulohumeral muscular dystrophy. *J. Med. Genet.* **33**, 366–370 (1996).
- Chaudhuri, T., Mukherjee, M., Sachdev, S., Randall, J. D. & Sarkar, S. Role of the fetal and α/β exons in the function of fast skeletal troponin T isoforms: correlation with altered Ca^{2+} regulation associated with development. *J. Mol. Biol.* **352**, 58–71 (2005).
- Johnson, J. M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
- Dubowitz, V. *Muscle Biopsy. A Practical Approach* (Bailliere Tindall, London, 1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to all FSHD families and the FSH Society. We thank M. Mora for providing FSHD myoblasts; G. Di Giulio for performing the radiology; C. Ghigna, S. Jones, L. Castilla and members of the M.R.G. laboratory for helpful discussions; and S. Evans for assistance with preparing the manuscript. Space constraints have limited the amount of original work that we have been able to cover in this Letter. This work was supported by the Muscular Dystrophy Association (D.G.); Associazione Amici del Centro Dino Ferrari, Telethon Bank, Eurobiobank and R.F. 2002 Criobanca Automatizzata di Materiale Biologico (M.M.); Telethon (R.B.); and the Cariplo Foundation, NIH-NINDS, the Muscular Dystrophy Association, the Italian National Council for Research Progetto Genomica Funzionale, the Association Francaise contre les Myopathies, the Italian Ministry of Education, University, and Research, Legato Ferrari and Telethon (R.T.). M.R.G. is an investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.T. (rossella.tupler@umassmed.edu).

LETTERS

Efficient auditory coding

Evan C. Smith^{1,2} & Michael S. Lewicki^{2,3}

The auditory neural code must serve a wide range of auditory tasks that require great sensitivity in time and frequency and be effective over the diverse array of sounds present in natural acoustic environments. It has been suggested^{1–5} that sensory systems might have evolved highly efficient coding strategies to maximize the information conveyed to the brain while minimizing the required energy and neural resources. Here we show that, for natural sounds, the complete acoustic waveform can be represented efficiently with a nonlinear model based on a population spike code. In this model, idealized spikes encode the precise temporal positions and magnitudes of underlying acoustic features. We find that when the features are optimized for coding either natural sounds or speech, they show striking similarities to time-domain cochlear filter estimates, have a frequency-bandwidth dependence similar to that of auditory nerve fibres, and yield significantly greater coding efficiency than conventional signal representations. These results indicate that the auditory code might approach an information theoretic optimum and that the acoustic structure of speech might be adapted to the coding capacity of the mammalian auditory system.

A fundamental issue in auditory coding is the nature of the computations that transform the raw sensory signal into a representation that is useful for auditory tasks. The response properties of cochlear nerves have been studied extensively, and models based on these results capture many properties of the neural response^{6,7}. However, these properties and auditory coding are still poorly understood in terms of underlying theoretical principles. Of the many sensory codes that could have existed, why has nature chosen one in particular, either through adaptation or evolution?

Traditional views describe auditory coding in terms of spectral features, such as frequency, intensity and phase, that are estimated from the signal. This perspective focuses on the properties and response of the system rather than its purpose. In contrast, theoretical approaches seek to predict properties of the system from underlying principles. What are these principles? One hypothesis is that of efficient coding, which posits that one of the primary goals of sensory coding is to form an efficient code, namely one that maximizes the amount of information conveyed about the sensory signal to the rest of the brain^{1–5}. In the peripheral auditory system, the incoming acoustic signal is transmitted mechanically to the inner ear and undergoes a highly complex transformation before it is encoded by spikes at the auditory nerve. If the auditory code is efficient, it should be possible to predict its properties from a theoretically ideal code.

To test this hypothesis, we must start with a mathematical description of an acoustic waveform that can then be used to derive theoretically optimal codes. Here we use a model in which sounds are encoded as a pattern of spikes^{8–10}. The signal, $x(t)$, is encoded with a set of kernel functions, $\phi_1, \dots, \phi_m$, that can be positioned arbitrarily and independently in time. The mathematical form of the representation with additive noise is

$$x(t) = \sum_i \sum_m s_i^m \phi_m(t - \tau_i^m) + \varepsilon(t) \quad (1)$$

where τ_i^m and s_i^m are the temporal position and coefficient of the i th instance of kernel ϕ_m , respectively. Note that the number of instances of ϕ_m need not be the same across kernels. To allow the kernel functions to assume arbitrary potential shapes, we represented each kernel ϕ_m by a vector of length L_m , where each element is an independent parameter of the model. Both the kernel shapes and their lengths were adapted to optimize coding efficiency; in the results below, the kernels take on a variety of shapes and range in length from ten to several hundred milliseconds. This provides a mathematical description of sound waveforms that has sufficient flexibility to encode arbitrary acoustic signals and encompass a broad range of potential auditory codes.

The key theoretical abstraction of the model is that the acoustic signal can be encoded most efficiently by decomposing it in terms of discrete acoustic elements, each of which has a precise amplitude and temporal position. This also yields a code that is time-relative and does not depend on artificial blocking of the signal¹⁰. One interpretation of each analogue τ_i^m, s_i^m pair is that it represents a local population of (binary) auditory nerve spikes firing probabilistically in proportion to the underlying analogue value. The form of the model allows for the case in which the coefficients s_i^m are constrained to be binary, but for computational reasons we have used analogue spikes as an approximation. To optimize the theoretical model to code natural sounds efficiently, we first need to address two problems: first, encoding (determining the optimal values of τ_i^m and s_i^m) and second, learning (determining the optimal kernel functions ϕ_m). From equation (1), coding efficiency can be defined approximately as the number of spikes required to achieve a desired level of precision, which is defined by the variance of the additive noise $\varepsilon(t)$. This assumes that the goal of coding is to represent the entire acoustic signal and that coding efficiency is most closely related to the number of spikes in the code. Other definitions are possible within this framework, but this definition has the advantage of starting from a minimal set of assumptions.

It is important to distinguish between the code and the encoding algorithm. For a given code (for example equation (1)), there are many different encoding algorithms that make different trade-offs in terms of coding efficiency, representational precision and computational complexity. Although the generative form of the model is linear, in other words the signal is a linear function of the representation, inferring the optimal representation for a signal is highly nonlinear and computationally complex. In fact, the problem of finding the optimal sparse representation by using a generic dictionary of functions is NP-hard¹¹, so only approximate algorithms are feasible. Here we compute the values of τ_i^m and s_i^m for a given signal by using a matching pursuit algorithm¹², which iteratively approximates the input signal and has been shown to yield highly efficient representations for a broad range of sounds¹⁰. Note also that although we have assumed a representation that consists of spikes (that is, a localized representation of the time position of an underlying acoustic feature¹³), spikes themselves are a consequence

¹Department of Psychology, ²Center for the Neural Basis of Cognition and ³Department of Computer Science, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213, USA.

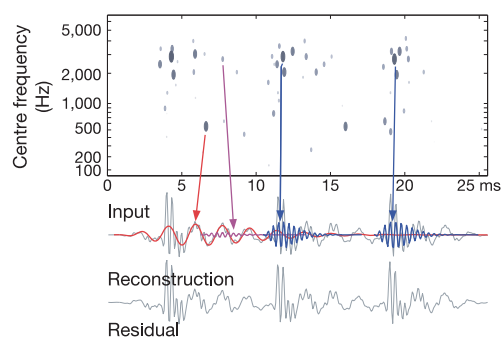


Figure 1 | Representing a natural sound with the use of spikes. A brief segment of the word ‘canteen’ (input) is represented as a spike code (top). Each spike (oval) represents the temporal position and centre frequency of an underlying kernel function, with oval size and grey value indicating kernel amplitude. The coloured arrows illustrate the correspondence between the spikes and the underlying acoustic structure represented by the kernel functions. Alignment of the spikes with respect to the kernels is arbitrary and is an issue only for plotting. We choose the kernel centre of mass, which for a delta-function input yields aligned spikes across the kernel population. A reconstruction of the speech from only the 60 spikes shown is accurate with little residual error (reconstruction and residual).

of optimizing the efficiency of the representation: if the coefficients s_i^m are assumed to be continuous in time and then optimized to represent the signal efficiently, only a discrete set of temporally sparse coefficients emerges^{8–10,14}.

Figure 1 illustrates the spike code model and its efficiency in representing speech. The spoken word ‘canteen’ was encoded with a set of spikes with the use of a fixed set of kernel functions (because the kernels can have arbitrary shape, for illustration purposes here we have chosen gammatones, mathematical approximations of cochlear filters). A brief segment from the input speech signal (Fig. 1, input) consists of three glottal pulses in the /a/ vowel. The resulting spike code is shown above it. The coloured arrows and curves indicate the relationship between the spikes (grey ovals) and the acoustic components they represent. The figure shows that a small set of spikes (for comparison, the sound segment contains about 400

samples) is sufficient to produce a very accurate reconstruction of the sound (Fig. 1, reconstruction and residual).

The spike-coding algorithm provides a way to encode signals given a set of kernel functions, but the actual efficiency of this code depends on how well the kernel functions capture the acoustic structure of the sound ensemble. To optimize the kernel functions we derived a gradient-based algorithm for adapting each kernel in shape and length to improve the fidelity of the representation (Supplementary Methods). Information theory states that there is a fundamental relationship between the efficiency of a code and the degree to which it captures the statistical structure of the signals being encoded. Thus, one of the primary tenets of efficient coding theory is that sensory codes should be adapted to the statistics of the relevant sensory environment. To make predictions, it is necessary to optimize the code to an ensemble of sounds to which the auditory system is thought to be adapted. However, this poses a problem because the precise composition of the natural acoustic environment is unknown, and many common sounds, such as wind noise, may have much less behavioural relevance than other sounds.

To address this issue, we made the generic assumption that the auditory system is adapted to an unknown mixture of three broad categories of natural sounds. The kernel functions were optimized to an ensemble of natural sounds that consisted of mammalian vocalizations¹⁵ and two subclasses of environmental sounds (Supplementary Methods). These sound classes represent a wide range of acoustic structure. Vocalizations tend to be harmonic and more steady-state, whereas environmental sounds have little or no harmonic structure and are more transient. Furthermore, to obtain an ensemble composition that yielded a good match to the physiological data (described below), we found it necessary to divide environmental sounds into two subclasses, namely transient environmental sounds, such as cracking twigs and crunching leaves, and ambient environmental sounds, such as rain and rustling sounds. This approach has the added advantage that we can investigate how the theoretically ideal code changes as a function of the sound ensemble composition.

Figure 2a shows the learned kernel functions (red curves) for the natural sounds ensemble. All kernels are time-localized, have a narrow spectral bandwidth and show a strong temporal asymmetry not predicted by previous theoretical models. The sharp attack and

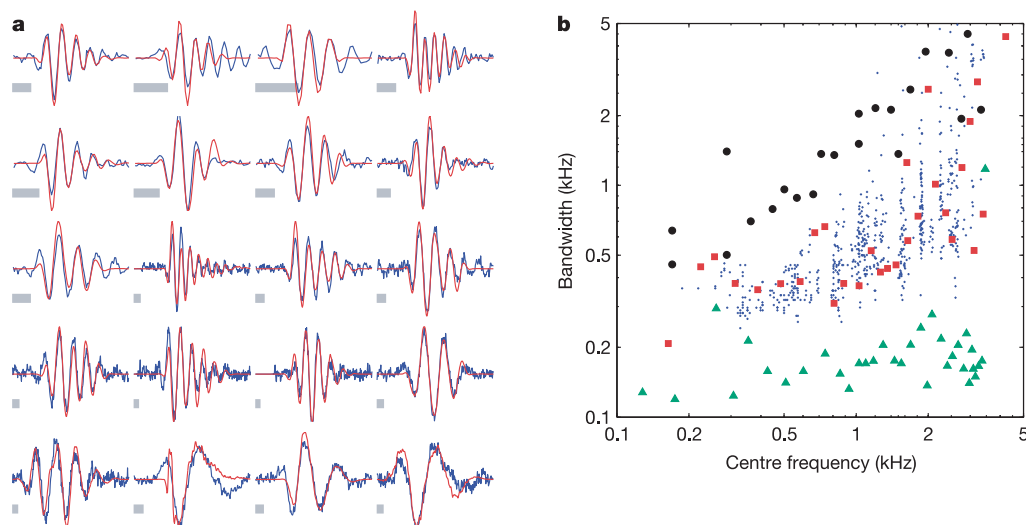


Figure 2 | Efficient codes for natural sounds predict revcor filter shapes and population characteristics. **a**, When optimized to encode an ensemble of natural sounds, kernel functions become asymmetric sinusoids (smooth curves in red, with padding removed) with sharp attacks and gradual decays. They also adapt in temporal extent, with longer and shorter functions emerging from the same initial length (grey scale bars, 5 ms). Each kernel

function is overlaid on a revcor function obtained from cat auditory nerve fibres (noisy curves in blue). **b**, The bandwidth–centre frequency distribution of learned kernel functions (red squares) is plotted together with cat physiological data (small blue dots) and with kernel functions trained on environmental sounds alone (black circles) or animal vocalizations alone (green triangles).

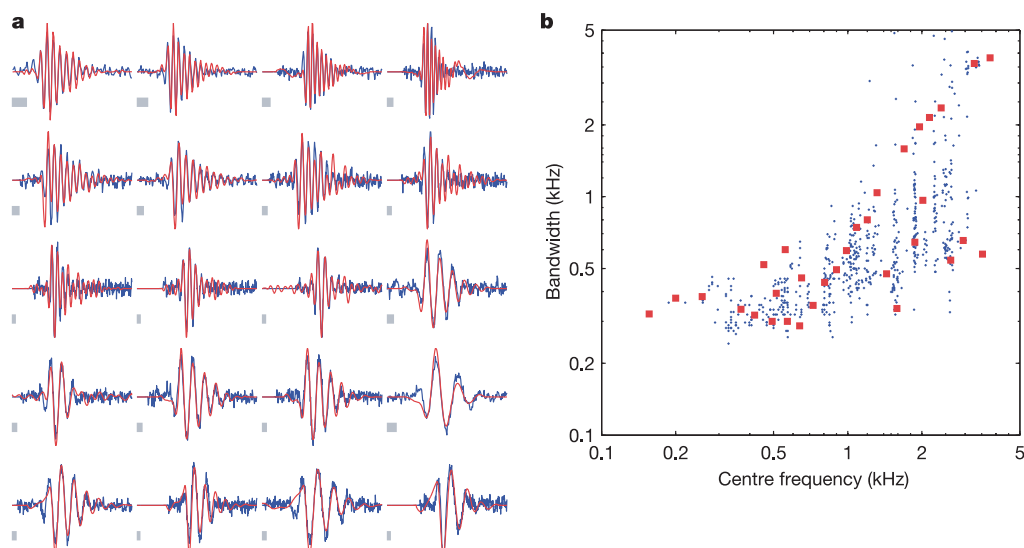


Figure 3 | Human speech is adapted to the mammalian cochlear code. **a**, As with the kernel functions trained on the natural sounds ensemble, the efficient code for speech consists of asymmetric sinusoids that closely match

auditory revcor filters. **b**, The population of speech-trained kernels also matches the population centre bandwidth–frequency relationship of cochlear revcor filters. Details are as in Fig. 2.

gradual decay of the envelope match the physiological filtering properties of auditory nerves as characterized by reverse-correlation, which estimates the impulse response function of individual auditory nerve fibres^{16–18}, the so-called ‘revcor’ filters. Kernel functions in Fig. 2a are overlaid on revcor filters obtained from cat auditory nerves¹⁹. The revcor functions were normalized, and the most similar (as determined by correlation) were aligned in phase with each of the learned kernels. Even though the learned kernel functions were derived entirely independently of the physiological revcor functions, they are strikingly similar. The similarity of the experimental data and theoretical predictions is not a result of selection bias. Statistical comparisons of the different models (both theoretical and parametric) show that the kernels derived theoretically have a slightly higher median correlation with the revcor filters than do parametric gammachirp models¹⁹ fitted to the same set of data (Supplementary Fig. S2). Repeated training from different random initial conditions produced similar results. We can also compare population properties of the learned kernels with those of the physiological data. Figure 2b shows a log–log scatter-plot of the bandwidth against centre frequency for each kernel (red squares) and for the physiological revcor data (small blue dots)²⁰. Both the slope and spread of the learned kernel functions match those of the empirical data, and the distribution seems to follow shifts in the slope and spread at low and high frequencies.

In contrast with the efficient code for the natural sounds ensemble, kernel functions optimized for other sound classes predict neither the structure of revcor filters nor their population characteristics (Fig. 2b). Efficient codes for animal vocalizations (green triangles) have kernels with much narrower bandwidths that do not increase with frequency. Kernels optimized for only environmental sounds (black circles) have bandwidths that increase much more rapidly with frequency. Furthermore, their respective kernel functions are significantly different from mammalian revcor filters (Supplementary Fig. S1). However, kernel functions optimized for speech predict both the asymmetric revcor structure and the population distribution just as well as the kernel functions optimized for the natural sounds ensemble (Fig. 3, and Supplementary Fig. S2).

We can quantify the coding efficiency of the learned kernel functions to evaluate the model objectively and compare it quantitatively with other signal representations. Fidelity–rate curves, which plot the fidelity of the encoded signal (the signal-to-noise ratio in decibels) against the coding rate in bits per second, provide a useful,

objective measure for comparison. Here we use a previously developed method¹⁰ in which we vary the precision of the spike amplitude and temporal position values, s_i^m and τ_i^m , and compute the resulting fidelity from the reconstruction error. This produces a curve showing the trade-off between quality and cost for a given representation. Fidelity–rate curves for speech coding were calculated for Fourier transform, for discrete wavelet transform and for spike codes by using either gammatones or kernel functions optimized for speech (Supplementary Methods). For all fidelity–rate curves (Fig. 4), increasing the information rate (x axis) always improves fidelity (y axis). Below 30 dB, the spike codes using both the optimized and fixed kernels (gammatones) resulted in more efficient representations of speech than the traditional representations. At a fidelity of 15 dB, the learned spike code is over threefold more efficient than either the Fourier or wavelet codes (8 kilobits s^{-1} versus 30 kilobits s^{-1}). Spike codes using learned kernels are also more

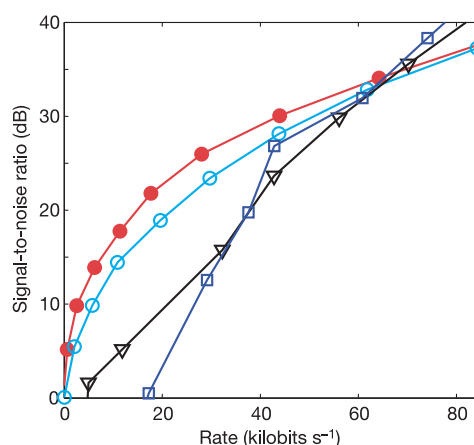


Figure 4 | Fidelity–rate curves for Fourier, wavelet and spike codes. The curves show the trade-off between coding cost and signal fidelity in the representation of speech signals from the TIMIT Speech Corpus testing set. Curves were generated for spike coding by using both speech-trained kernels (red circles) and gammatones (light blue circles) as well as with a discrete Daubechies wavelet transform (dark blue squares) and a Fourier transform (black triangles). Confidence intervals are extremely tight (see text) and are not plotted.

efficient than those using fixed gammatones. At fidelities above 35 dB, the residual signal being coded is not significantly different from gaussian noise¹⁰ and is described more efficiently with the Fourier or wavelet representation.

These results provide a theoretical explanation of the shape of auditory revcor filters. Previous explanations of the revcor filter shapes have been phenomenological, being a consequence of the impulse response function of the basilar membrane. Our results show that the rapid rise and slow decay that is characteristic of the revcor envelope is ideally matched to the statistical structure of natural sounds and indicate that it might not be an arbitrary feature of the biological system. In this sense, the physics of basilar membrane response (and subsequent cochlear processing) is tuned to the vibrations of objects in the natural environment. Rapid onsets followed by slower decays are characteristic of many types of transient natural sound and provide acoustic cues for essential auditory tasks such as sound localization. Note, however, that rapid rise and slow decay are not present in all efficient codes or an artefact of the learning algorithm, because the optimal kernels for vocalizations are largely symmetric and kernels adapted to reverse speech show the opposite pattern: slow rise and rapid decay (Supplementary Fig. S1).

Filterbank models of the cochlear have assumed that for a given frequency there is a fixed filter bandwidth (corresponding to gammatone function of fixed length). The auditory revcor data show a range of bandwidths at any given frequency, which is also matched by the kernel functions derived theoretically. One interpretation of this finding is that to form an efficient code for a variety of sounds it is necessary to have kernel functions of varying lengths to provide a better description of sounds with different temporal correlation constants.

We have compared the learned kernels with the physiological revcor filters, but it is not obvious why the comparison should be appropriate. Reverse correlation provides a first-order characterization of a nonlinear system, but there is no direct way to measure the 'features' that the auditory code is using to describe the acoustic signal. However, the use of reverse correlation to derive the equivalent revcor filters from the theoretical model yields an almost exact match to the learned kernel functions (Supplementary Fig. S3). For reverse correlation to recover the auditory nerve impulse response functions, neural spikes must be uncorrelated, which is often assumed to arise from stochastic firing. In contrast, spikes in the theoretical model are uncorrelated because they precisely encode non-redundant information. This indicates that if neural spikes are uncorrelated for the same reason, the revcor 'filters' might be more analogous to the acoustic features of the auditory code.

Our results depended on obtaining spike representations that are efficient. Learning experiments with an encoding algorithm that yields less efficient representations but is more biologically plausible (one in which the acoustic waveform was convolved with each of the kernel functions and thresholded to obtain the spike trains) produced kernel functions that showed no similarity to the auditory revcor functions (E.C.S and M.S.L., unpublished observations).

Although these results offer an explanation of auditory revcor data, we caution that there remain several challenges in extending the scope of the theory to auditory nerve responses. Algorithms for optimally encoding sound with binary spikes would permit a more direct comparison with auditory nerve firing patterns. A more accurate description of the encoding objective is also possible. Exact representation of the sound pressure waveform, as assumed here, is an unlikely goal, because not all acoustic information is behaviourally relevant. The gradual decrease phase-locking properties for higher-frequency nerve fibres might provide one indication of what information the auditory system encodes. There is also no explanation for the role of adaptation or changes in response with stimulus intensity, which are probably important in maximizing the information conveyed through the auditory nerve.

Finally, we note that deriving efficient codes for speech immediately yielded kernels that closely matched the auditory revcor filters. In contrast, obtaining similar results for the natural sounds ensemble required careful balancing of the different sound categories. This indicates that the acoustic composition of speech itself might be adapted to the mammalian auditory system.

METHODS

Encoding. Signals were encoded with a matching pursuit-based algorithm^{10,12}, which iteratively decomposed the signal in terms of the kernel functions. The current residual signal (or the original signal) was projected onto the dictionary of kernel functions. The projection with the largest inner product was subtracted out, and its coefficient and time were recorded. For the results reported here, the encoding was halted when s_i^m fell below a preset 'spiking' threshold. Further details are given in Supplementary Methods.

Learning. Equation (1) in the main text can be rewritten in probabilistic form in which we assume that the noise is gaussian and the prior provability of a spike, $p(s)$, is sparse. The kernel functions are optimized by performing gradient ascent on the approximate log data probability,

$$\begin{aligned} \frac{\partial}{\partial \phi_m} \log(p(x|\Phi)) &= \frac{\partial}{\partial \phi_m} \log(p(x|\Phi, \hat{s})) + \log(p(\hat{s})) \\ &= \frac{1}{2\sigma_e} \frac{\partial}{\partial \phi_m} \left[x - \sum_{m=1}^M \sum_{i=1}^{n_m} \hat{s}_i^m [x - \hat{x}]_{\tau_i^m} \right]^2 \\ &= \frac{1}{\sigma_e} \sum_i \hat{s}_i^m [x - \hat{x}]_{\tau_i^m} \end{aligned} \quad (2)$$

where $[x - \hat{x}]_{\tau_i^m}$ indicates the residual error over the extent of kernel ϕ_m at position τ_i^m . The estimated kernel gradient is thus a weighted average of the residual error. For training, 32 kernel functions were initialized as 100-sample gaussian noise, and the spiking threshold (minimum value of s_i^m) was set at 0.1. Further details are given in Supplementary Methods.

Sounds. The natural sounds ensemble used in training combined a collection of mammalian vocalizations with two classes of environmental sounds recorded by the authors in both natural and anechoic settings: ambient sounds (rustling brush, wind, flowing water) and transients (snapping twigs, crunching leaves, impacts of stone or wood). The reported 'natural sounds' results were based on (encoded) power proportions of 1.0:0.8:1.2, respectively, which directly reflects the impact of each sound class on kernel function adaptation. Speech was obtained from the TIMIT continuous speech corpus. All sounds were converted to a sampling frequency of 16 kHz, bandpass filtered to be between 100 and 6,000 Hz, and normalized to have a maximum amplitude of 1. Spike encodings of 5–40 s of training sounds were used to estimate the gradients on each update of the kernel functions.

Received 18 June; accepted 30 November 2005.

- Barlow, H. B. in *Sensory Communication* (ed. Rosenbluth, W. A.) 217–234 (MIT Press, Cambridge, Massachusetts, 1961).
- Atick, J. J. Could information-theory provide an ecological theory of sensory processing. *Network* **3**, 213–251 (1992).
- Simoncelli, E. & Olshausen, B. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
- Laughlin, S. B. & Sejnowski, T. J. Communication in neuronal networks. *Science* **301**, 1870–1874 (2003).
- Lewicki, M. S. Efficient coding of natural sounds. *Nature Neurosci.* **5**, 356–363 (2002).
- Shamma, S., Chadwick, R., Wilbur, W. J., Morrish, K. A. & Rintel, J. A biophysical model of cochlear processing: Intensity dependence of pure tone responses. *J. Acoust. Soc. Am.* **78**, 1612–1621 (1986).
- Yang, X., Wang, K. & Shamma, S. Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory* **38**, 824–839 (1992).
- Lewicki, M. S. & Sejnowski, T. J. in *Advances in Neural Information Processing Systems* (eds Kearns, M. J., Solla, S. A. & Cohn, D. A.) vol. 11, 730–736 (MIT Press, Cambridge, Massachusetts, 1999).
- Lewicki, M. S. in *Probabilistic Models of the Brain: Perception and Neural Function* (eds Rao, R. P. N., Olshausen, B. A. & Lewicki, M. S.) 241–255 (MIT Press, Cambridge, Massachusetts, 2002).
- Smith, E. C. & Lewicki, M. S. Efficient coding of time-relative structure using spikes. *Neural Comput.* **17**, 19–45 (2005).
- Davis, G., Mallat, S. & Avellaneda, M. Adaptive greedy approximations. *Construct. Approx.* **13**, 57–98 (1997).
- Mallat, S. G. & Zhang, Z. Matching pursuits with time–frequency dictionaries. *IEEE Trans. Signal Process.* **41**, 3397–3415 (1993).
- de Ruyter van Steveninck, R. & Bialek, W. Realtime performance of a

- movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proc. R. Soc. Lond. B* **234**, 379–414 (1988).
14. Olshausen, B. A. in *Probabilistic Models of the Brain: Perception and Neural Function* (eds Rao, R. P. N., Olshausen, B. A. & Lewicki, M. S.) 257–272 (MIT Press, Cambridge, Massachusetts, 2002).
 15. Emmons, L. H., Whitney, B. M. & Ross, D. L. *Sounds of the Neotropical Rainforest Mammals* [audio CD] (Library of Natural Sounds, Cornell Laboratory of Ornithology, Ithaca, New York, 1997).
 16. deBoer, E. & deJongh, H. On cochlear encoding: Potentialities and limitations of the reverse correlation technique. *J. Acoust. Soc. Am.* **63**, 115–135 (1978).
 17. Carney, L. H. Sensitivities of cells in the anteroventral cochlear nucleus of cat to spatiotemporal discharge patterns across primary afferents. *J. Neurophysiol.* **64**, 437–456 (1990).
 18. Recio-Spinoso, A., Temchin, A. N., van Dijk, P., Fan, Y.-H. & Ruggero, M. A. Wiener-kernel analysis of responses to noise of chinchilla auditory-nerve fibers. *J. Neurophysiol.* **93**, 3635–3648 (2005).
 19. Irino, T. & Patterson, R. A level-dependent auditory filter: the gammachirp. *J. Acoust. Soc. Am.* **101**, 764–774 (1997).
 20. Carney, L. H., McDuffy, M. J. & Shekhter, I. Frequency glides in the impulse responses of auditory-nerve fibers. *J. Acoust. Soc. Am.* **105**, 2384–2391 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements E.C.S. was supported by a National Institutes of Health training grant. This material is based on work supported by a National Science Foundation grant to M.S.L. Empirical data in this paper were acquired from Boston University's Earlab, an online, freely accessible auditory database (<http://earlab.bu.edu>).

Author Contributions M.S.L. and E.C.S. developed the model, analysed the results and wrote the paper together; E.C.S. designed and ran the simulations.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S.L. (lewicki@cnbc.cmu.edu).

Fast vesicle reloading and a large pool sustain high bandwidth transmission at a central synapse

Chiara Saviane¹ & R. Angus Silver¹

What limits the rate at which sensory information can be transmitted across synaptic connections in the brain? High-frequency signalling is restricted to brief bursts at many central excitatory synapses^{1,2}, whereas graded ribbon-type synapses can sustain release³ and transmit information⁴ at high rates. Here we investigate transmission at the cerebellar mossy fibre terminal, which can fire at over 200 Hz for sustained periods *in vivo*⁵, yet makes few synaptic contacts onto individual granule cells⁶. We show that connections between mossy fibres and granule cells can sustain high-frequency signalling at physiological temperature. We use fluctuation analysis⁷ and pharmacological block of desensitization to identify the quantal determinants of short-term plasticity and combine these with a short-term plasticity model and cumulative excitatory postsynaptic current analysis to quantify the determinants of sustained high-frequency transmission. We show that release is maintained at each release site by rapid reloading of release-ready vesicles from an unusually large releasable pool of vesicles⁸ (~300 per site). Our results establish that sustained vesicular release at high rates is not restricted to graded ribbon-type synapses and that mossy fibres are well suited for transmitting broad-bandwidth rate-coded information to the input layer of the cerebellar cortex.

Cerebellar granule cells (GCs), the smallest neurons in the brain, receive an average of four mossy fibre (MF) inputs, one onto each of their short dendrites. The excellent voltage-clamp afforded by the GC and the fact that only one of the MF inputs is likely to be activated by local stimulation in acute slices⁹ make this central synapse an attractive model system for studying neurotransmitter release. *In vivo* studies show that MFs encode various sensory modalities, including transient whisker deflection¹⁰ and continuous signalling of joint angle⁵, and that GCs can fire up to 250 Hz without accommodating¹⁰. Here we have focused on how MFs transmit sustained broad-bandwidth rate-coded signals⁵ to GCs. We have examined the frequency dependence of AMPAR (α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor)-mediated transmission at MF–GC connections by applying trains of 20–100 stimuli at 20–300 Hz (Fig. 1), covering the sustained firing rates observed *in vivo*⁵. As the frequency was increased, the phasic excitatory postsynaptic current (EPSC), which is related to direct release onto the GC, was accompanied by a tonic component, part of which is probably due to glutamate spillover from release sites onto neighbouring GCs, a major determinant of the EPSC decay time course¹¹ (Fig. 1a–c). The onset of depression of the phasic EPSC was rapid, reaching a steady-state level of $34 \pm 3\%$ of the initial amplitude during 100-Hz trains (double-exponential fit with time constants of $\tau_1 = 1.5$ pulses (86%) and $\tau_2 = 25$ pulses; Fig. 1d, e). The phasic component of the EPSC declined more strongly with frequency than the total current (Fig. 1e), suggesting that at high frequencies much of the current is due to summation of the slow current components of the EPSCs. The

total steady-state charge transfer per second, normalized by the first EPSC, increased up to 300 Hz (Fig. 1f), suggesting that MFs can signal to GCs over a wide range of frequencies. The shallow frequency dependence of the phasic component contrasts with many depressing excitatory central synapses in mature preparations^{12,13}, but is comparable to that observed at calyx of Held synapses^{14,15}, which have hundreds of release sites¹⁶. Our results show that, despite few synaptic contacts, MF–GC synapses can sustain release over a wide range of frequencies, suggesting that replenishment of release-ready vesicles (RRVs) is rapid.

We investigated the mechanisms underlying MF–GC transmission

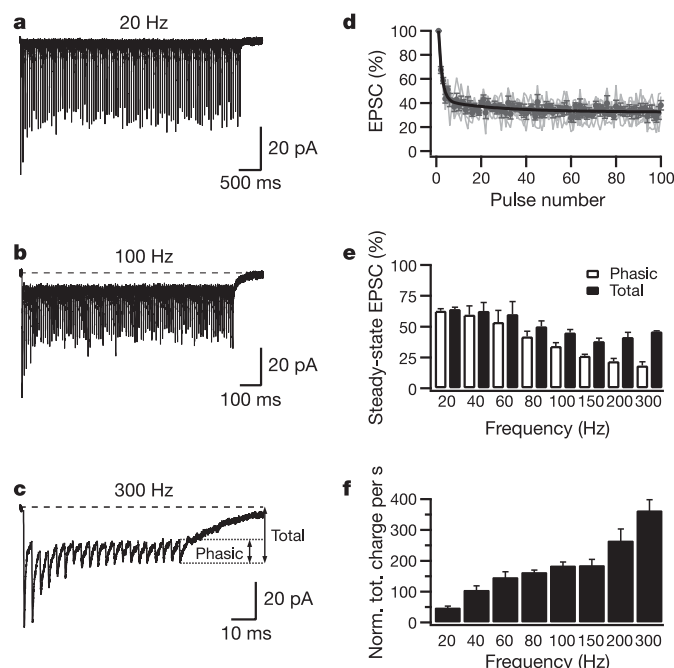


Figure 1 | Frequency dependence of transmission at the MF–GC synapse. **a, b**, Mean EPSCs during 100-pulse trains at 20 and 100 Hz, respectively (stimulus artefacts are blanked for clarity). **c**, Mean EPSCs during a 20-pulse train at 300 Hz (different cell). **d**, Normalized phasic EPSC amplitude during 100-Hz trains for five MF connections together with the average (grey dots) and double-exponential fit (black line). **e**, Mean steady-state EPSC amplitude for phasic and total current, normalized to the first EPSC ($n = 4-9$). Train length was 100 pulses at 20–100 Hz and 20 pulses at 150–300 Hz (≥ 7 trains; intersweep interval, 15 s). **f**, Total steady-state charge per second, calculated from the mean area per pulse (over the last five pulses) and normalized by the first EPSC amplitude and the pulse duration.

¹Department of Physiology, University College London, Gower Street, London WC1E 6BT, UK.

by examining quantal parameters during stimulation trains using a combination of fluctuation analysis methods (Supplementary Information). Because most of the depression occurred in the first few stimuli (Fig. 1d), we used a reduced protocol of five pulses at 100 Hz (Fig. 2a). We separated slow-rising spillover-mediated EPSCs from fast-rising currents¹¹ and used the spillover current to isolate the quantal component of the EPSC¹⁷ (Fig. 2b and Methods). Quantal parameters underlying the first EPSC were determined for each connection from the spillover-corrected mean and variance¹⁷ measured in different concentrations of Ca^{2+} , which imposed different release probabilities, by fitting a multinomial model of release to the variance–mean relationship (multiple-probability fluctuation analysis (MPFA)⁷; Fig. 2c, d). Because quantal sum linearly at this synapse¹⁷, MPFA provided the mean number of functional release sites per connection (N , 5.3 ± 0.8 , mean $\pm$ s.e.m.; $n = 9$).

The mean quantal size at the peak of the quantal component (Q_P) of the first EPSC was not significantly different from that measured directly under low release probability conditions¹⁷ ($Q_P = -14.9 \pm 1.4$ pA; unpaired t -test, $P = 0.61$), and the mean probability of release per site (P_R , 0.61 ± 0.07 in 2 mM Ca^{2+}) was in line with previous estimates^{17,18}. Knowing N for each connection, we were then able to determine P_R and Q_P for the second and fifth pulse, using spillover-corrected coefficient of variation analysis (Fig. 2e, f). The nine MF connections analysed had a wide range of initial P_R values with both depression and facilitation in P_R being observed during the train, whereas Q_P was consistently depressed. On average, P_R did not change significantly on the second ($89 \pm 7\%$) or fifth ($81 \pm 10\%$) pulse, indicating that P_R is maintained at an intermediate level during 100-Hz trains (Fig. 2g). By contrast, Q_P was significantly reduced on the second and fifth pulse ($81 \pm 5\%$ and $69 \pm 3\%$, respectively; one-way analysis of variance (ANOVA), $P < 0.0001$; Fig. 2g).

Application of cyclothiazide (CTZ) to reduce desensitization and kynurenic acid to prevent receptor saturation¹⁹ diminished the amount of depression ($n = 6$; two-way ANOVA, $P < 0.05$) to a level comparable to the mean depression in P_R (Supplementary Fig. 1), indicating that depression in Q_P was due to AMPAR desensitization. Although several presynaptic mechanisms could

underlie the depression in P_R (ref. 16), the strong negative correlations between the change in P_R from the first to the second pulse ($r = -0.92$, $P < 0.001$, Fig. 2h) or from the first to the fifth pulse ($r = -0.87$, $P < 0.01$) and its initial value indicate that the depression in P_R is caused by a release-dependent mechanism, which is often assumed to reflect vesicle depletion or, more generally, inactivation of the release site after release²⁰. The presence of facilitation and depression in P_R at different fibres suggests that changes in P_R during the train are set by two factors²¹: the fraction of release sites with a RRV; and the release probability of RRVs, which increases owing to facilitation.

To understand how P_R is maintained during high-frequency trains, we examined the kinetics of the underlying processes with paired-pulse protocols. Under control conditions the relationship between the mean paired-pulse ratio (PPR, ratio of the mean amplitude of the second and first EPSC) and the interpulse interval (IPI) could be fit with two exponential components (with time constants and amplitudes of $\tau_1 = 12$ ms, $\tau_2 = 1,600$ ms and $A_1 = 0.27$, $A_2 = 0.09$, respectively; $n = 32$; Fig. 3a). Application of CTZ and kynurenic acid (Fig. 3b) significantly increased the PPR (two-way ANOVA, $P < 0.0001$), suggesting that even at early times most of the depression was mediated by AMPAR desensitization, consistent with our results from MPFA and with findings in younger rats²². Because it is difficult to separate the amplitudes and time courses of facilitation and depression under control conditions, owing to contamination by the spillover current and, possibly, spillover-mediated desensitization²², we isolated facilitation using a paired-pulse protocol under low P_R conditions (1.25–1.5 mM Ca^{2+} ; Fig. 3c) when the spillover current is minimal. By comparing the mean amplitude of the first EPSC with that of the second EPSC when no release occurred on the first stimulus, it was possible to quantify facilitation in the absence of presynaptic depression and postsynaptic desensitization. The time course of the mean facilitation component from 7 cells was fitted by a single-exponential function with a decay time constant (τ_F) of 12 ms (Fig. 3c, bottom).

We then estimated the time course of recovery from presynaptic depression using a simple short-term plasticity model that included facilitation and depression²¹ (Supplementary Information). We modelled release at each fibre using the individual measured initial

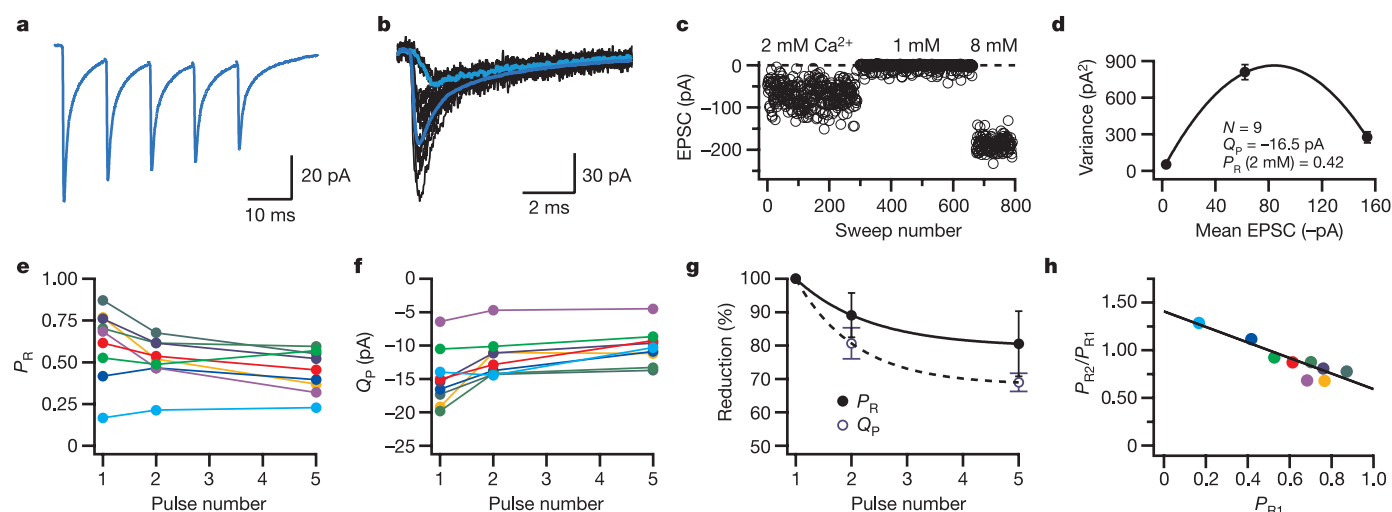


Figure 2 | Quantal determinants of short-term plasticity at MF-GC connections. **a**, Mean EPSCs during a five-pulse 100-Hz train. **b**, Individual slow-rising spillover-mediated and fast-rising EPSCs, together with their means (cyan and blue line, respectively), recorded on the first pulse. **c**, **d**, EPSC amplitudes in different concentrations of Ca^{2+} and associated spillover-corrected variance–mean plot, respectively. Line shows fit with multinomial model (Supplementary Information). Errors were calculated as described in ref. 7. **e**, **f**, Respective P_R and Q_P during five-pulse 100-Hz trains

in 2 mM Ca^{2+} for nine MF-GC connections. At least 145 consecutive traces were used for coefficient of variation analysis to minimize sampling error. **g**, Mean relative changes in P_R and Q_P during the train with associated mono-exponential fits (lines; P_R : $A = 20\%$ $\tau = 1.3$ pulses; Q_P : $A = 32\%$, $\tau = 1.1$ pulses). **h**, Relationship between the ratio of P_R on the second and first pulse and the initial P_R . The same colours as used in **e** and **f** are used for each cell.

P_R and the mean time course of facilitation, leaving the time constant of recovery from presynaptic depression (τ_R) as the only free parameter. The measured P_R during the five-pulse train was fitted to the model for each MF connection studied (Fig. 4a). Although the initial value of P_R and the subsequent presynaptic plasticity varied widely across connections, the estimate of τ_R was consistent and not correlated with P_R ($r = 0.28$, $P = 0.47$). The mean value of τ_R from the individual fits was 12 ± 2 ms ($n = 9$), similar to that calculated from the mean P_R across cells (13 ms). Figure 4b shows the mean time course of P_R through the train, together with its fit to the model and the underlying components: namely, the probability of a site being occupied and the release probability of RRVs. Because our approach assumes a specific facilitation model, with a magnitude that scales with P_R but a fixed decay time constant, we examined a range of τ_F values and other facilitation models that allowed the amplitude and/or the kinetics to change. These gave estimates of $\tau_R < 17$ ms, suggesting that it is relatively insensitive to the properties of facilitation. Moreover, the paired-pulse reduction in P_R predicted from our model for an initial P_R of 1 is in good agreement with the value predicted from Fig. 2h (57% and 59%, respectively at 10 ms). The rapid time course of recovery from depression allows RRVs to be replenished rapidly, and the similarity of the time course to that of facilitation allows a relatively constant vesicular output to be maintained during the train (Fig. 4b).

To understand how high-frequency transmission can be sustained for longer periods, we examined the size of the releasable pool of vesicles and its recovery rate by applying two 100-Hz trains of 6,000 stimuli with a 10-s interval. Figure 4c shows mean phasic EPSC amplitudes for the first and second train at connections where the pool was depleted during the first 6,000 stimuli (5 of 7 cells;

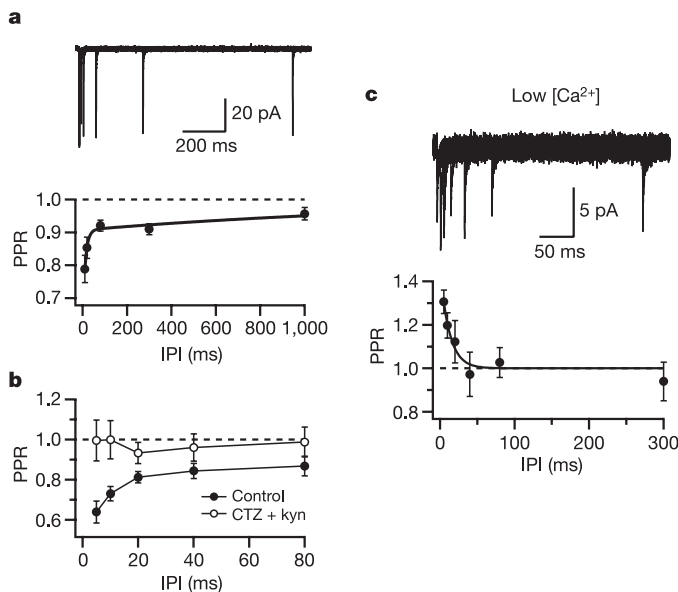


Figure 3 | Recovery from depression and facilitation at MF–GC synapses. **a**, Top, mean EPSCs evoked using a paired-pulse protocol with intervals of 10–1,000 ms. Bottom, relationship between average PPR and IPI for 32 MF–GC connections and double-exponential fit (solid line; asymptote constrained to 1). **b**, Mean PPR for shorter IPI (5–80 ms) in control and in CTZ and kynurenic acid (kyn) ($n = 7$). At least seven traces per IPI were recorded in **a** and **b**. **c**, Top, mean of the first EPSC together with the means of the second EPSCs obtained at the different IPIs (5–300 ms) when no release occurred on the first stimulus in 1.5 mM Ca^{2+} . Bottom, average PPR, calculated as in the top panel, versus IPI for 1.25–1.5 mM Ca^{2+} ($n = 7$; $P = 0.001$ and $P < 0.05$ at 5 and 10 ms, respectively) and single-exponential fit (line; asymptote constrained to 1). Cells were included if the spillover current was $< 7\%$ of the EPSC successes (≥ 15 traces per IPI).

Methods). We defined the releasable pool size as the number of vesicles released before the EPSC amplitudes reached steady state²³. The mean number of releasable vesicles per site was calculated by normalizing the back-extrapolated mean cumulative EPSC amplitude²³ by the mean NQ_P and correcting for desensitization (300; Fig. 4d and Methods). This represents roughly 40% of the total number of vesicles present per site (~ 700 from the 200,000 vesicles per MF terminal^{6,24} and ~ 300 release sites; that is, 5 onto each of ~ 53 GCs⁶). We then examined the rate at which the releasable pool was replenished following depletion. The amplitude of the first EPSC was not significantly different for the two trains, indicating that P_R had recovered in 10 s, but depletion of the pool occurred more rapidly during the second train, indicating that the pool had recovered to only 27% of the control value (Fig. 4c, d). This implies that the releasable pool is refilled at eight vesicles per second, a value identical to the steady-state release rate that we obtained from the slope of the cumulative amplitude after pool depletion.

Our results show that release sites at cerebellar MFs have an unusually large pool of releasable vesicles that is sevenfold larger than the ‘maxipool’ observed at hippocampal MFs²⁵. Refilling of the pool could arise from replenishment from a ‘reserve pool’ and/or rapid vesicle cycling through ‘kiss-and-run’ or fast endocytosis²⁶, as for hippocampal MFs²⁵. Achieving sustained high-frequency trans-

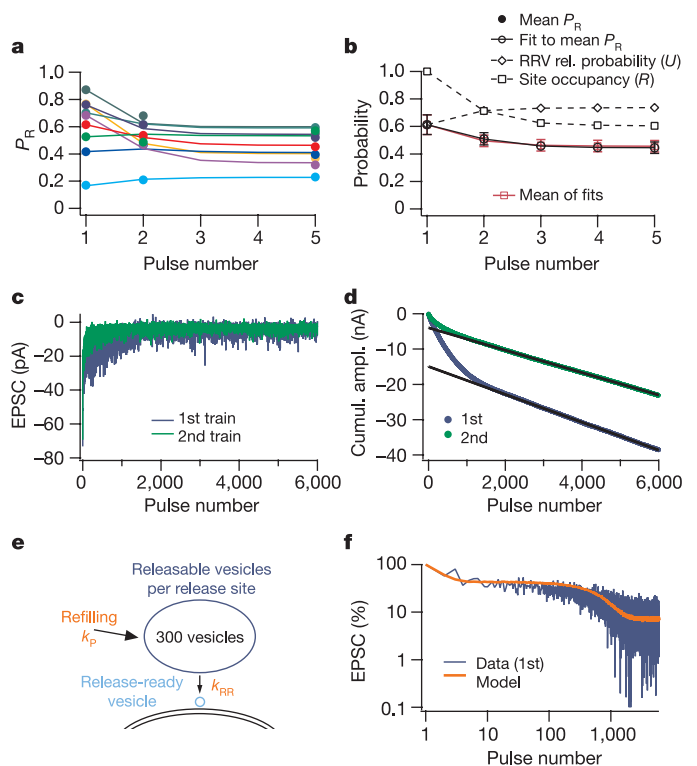


Figure 4 | Time course of vesicle reloading and releasable pool size. **a**, Fits of a plasticity model (lines) to measured values of P_R for nine MF–GC connections from Fig. 2, where τ_R was the only free parameter. **b**, Mean P_R across the nine cells is shown with the corresponding fit and the underlying facilitation of the probability of release of RRVs (U) and depression in the occupancy of release sites (R ; Supplementary Information). The fit of the mean is compared with the average of the fits from **a**. **c**, **d**, Mean peak phasic EPSC amplitude (**c**) and corresponding cumulative amplitude (**d**) during two trains of 6,000 stimuli at 100 Hz with a 10-s interval ($n = 5$). Lines represent back-extrapolated linear fits obtained from the last 3,000 pulses. **e**, Representation of the releasable pool and the RRV at each MF release site with the associated transition rates. **f**, Normalized mean EPSC amplitude during the 6,000-pulse train and fractional EPSC for the model in **e** plus receptor desensitization (5,000 repetitions).

mission at connections with few release sites through rapid reloading of RRVs from a large releasable pool (Fig. 4e) contrasts with the calyx of Held, where a small pool per site (<10) and the slow rate of vesicle cycling is circumvented by a low probability of release^{14,15} and 600 release sites¹⁶. The MF synaptic contact is one of the smallest in the brain (160-nm diameter²⁷) with an area threefold smaller than those at climbing and parallel fibres, which have 7–8 docked vesicles²⁸. Because the 300 vesicles in the releasable pool could not all be docked, vesicle translocation, docking and priming (TDP) must occur during sustained activity. Our simple model of transmission, based on our measurements (Fig. 4e plus desensitization; Supplementary Information), can predict reasonably well the fractional EPSC amplitude during 100-Hz stimulation over a wide timescale (Fig. 4f). This suggests that TDP can occur as fast as 12 ms during high-frequency trains, which is more than an order of magnitude faster than reported at the goldfish bipolar cell (250 ms; ref. 29) but comparable to the hair cell, where a replenishment rate of $1.1 \text{ vesicles ms}^{-1}$ is achieved through parallel processing of 14 docked vesicles per ribbon³. If the RRV at each site is replenished from docked vesicles that are refilled in parallel from the releasable pool, TDP could be slower. We modelled such a parallel system for the limiting case of four docked vesicles plus a RRV, which would correspond to a vesicle density twice that at climbing and parallel fibre synaptic contacts. This provided an upper estimate of 50 ms for the replenishment of docked, but not release-ready, vesicles. These limits indicate that TDP can be substantially faster than previously thought.

Our results show that sustained vesicular output is not limited by the number of docked vesicles and that the functional differences between central synapses and graded ribbon-type synapses are smaller than has been thought. The ability of MFs to sustain high-frequency transmission across few release sites is likely to be important for coordinating movements by permitting continuous rate-coded sensory information, such as joint angle⁵, to be transmitted to many GCs rapidly and with high fidelity.

METHODS

Parasagittal cerebellar slices were prepared from Sprague–Dawley rats aged 25–30 d (refs 9, 17). The external solution contained (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH_2PO_4 , 26 NaHCO_3 , 25 glucose, 2 CaCl_2 and 1 MgCl_2 (pH 7.4 when bubbled with 95% O_2 and 5% CO_2). Whole-cell recordings were made with pipettes containing (in mM): 105 KMeSO₄, 40 HEPES, 4 NaCl, 5 EGTA, 1.78 CaCl_2 , 0.3 NaGTP and 4 MgATP (pH 7.3) with a resistance of 6–10 M Ω . Recordings were made at 35.5–37.5 °C in the presence of 10 μM AP5, 20 μM 7-chlorokynurenic acid, 10 μM SR 95531 and 0.3 μM strychnine at -70 mV to isolate AMPAR-mediated current. Kynurenic acid (1.5–2 mM) and a low concentration of CTZ (25 μM) were sometimes added. This did not change the failure rate, ruling out presynaptic effects. Recordings were discarded if access resistance was $>35 \text{ M}\Omega$. Single MF inputs were evoked with a second pipette in the GC layer⁹. Recordings were made with an Axopatch 200B, filtered to 7.1 kHz and digitized at 50–100 kHz.

Data were analysed with Neuromatic software (http://www.physiol.ucl.ac.uk/research/silver_a/software). Stimulus artefacts were removed by subtracting a double-exponential fit to the decay of the artefact current. Spillover-corrected MPFA was done as described¹⁷ (Supplementary Information), but with a reduced protocol with three solutions containing different concentrations of Ca^{2+} and Mg^{2+} (in mM: 1/5, 2/1 and 8/0). For coefficient of variation analysis (Supplementary Information) and paired-pulse recovery, EPSC amplitudes were measured as described for MPFA¹⁷, after subtracting the double-exponential fit of the preceding EPSC. To achieve this we limited fluctuations analysis to the first, second and fifth pulse, and alternated five-pulse trains with single pulses and four-pulse trains. During longer trains (20–6,000 pulses), the peak amplitude was detected within a window of 1–2 ms after each artefact and was measured relative to a baseline of 0.1–0.5 ms preceding the pulse.

During the 6,000-pulse trains, stimulation intensity was raised to 20 V above threshold to counteract changes in stimulation threshold. We tested that this did not recruit more fibres by comparing EPSC amplitude to that near threshold. The pool was considered depleted if the steady-state amplitude reached the same asymptotic value during both trains. Desensitization was calculated at the end of

the train ($40 \pm 13\%$; $n = 4$) from the amplitude of spontaneous EPSCs, detected at least 4 ms after the artefact using a threshold criterion, and was not significantly different from the reduction in Q_P observed during the five-pulse train (unpaired t -test, $P = 0.35$; Fig. 2g), suggesting that desensitization reached a steady-state early in the train. Assaying vesicular release from the EPSC amplitude rather than from deconvolution³⁰ is unlikely to introduce significant error because vesicular release is highly synchronized at the MF¹⁷. Fits were weighted when errors on the data were known⁷. Sample means are expressed $\pm \text{s.e.m.}$ and were compared using the two-tailed paired Student's t -test, unless otherwise stated. Correlation between samples was tested with the Pearson's coefficient (r).

Received 6 October; accepted 25 November 2005.

1. Silver, R. A., Lubke, J., Sakmann, B. & Feldmeyer, D. High-probability unquantal transmission at excitatory synapses in barrel cortex. *Science* **302**, 1981–1984 (2003).
2. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl Acad. Sci. USA* **94**, 719–723 (1997).
3. Griesinger, C. B., Richards, C. D. & Ashmore, J. F. Fast vesicle replenishment allows indefatigable signalling at the first auditory synapse. *Nature* **435**, 212–215 (2005).
4. de Ruyter van Steveninck, R. R. & Laughlin, S. B. The rate of information transfer at graded-potential synapses. *Nature* **379**, 642–664 (1996).
5. van Kan, P. L., Gibson, A. R. & Houk, J. C. Movement-related inputs to intermediate cerebellum of the monkey. *J. Neurophysiol.* **69**, 74–94 (1993).
6. Jakab, R. L. & Hamori, J. Quantitative morphology and synaptology of cerebellar glomeruli in the rat. *Anat. Embryol.* **179**, 81–88 (1988).
7. Silver, R. A. Estimation of nonuniform quantal parameters with multiple-probability fluctuation analysis: theory, application and limitations. *J. Neurosci. Methods* **130**, 127–141 (2003).
8. Rizzoli, S. O. & Betz, W. J. Synaptic vesicle pools. *Nature Rev. Neurosci.* **6**, 57–69 (2005).
9. Silver, R. A., Cull-Candy, S. G. & Takahashi, T. Non-NMDA glutamate receptor occupancy and open probability at a rat cerebellar synapse with single and multiple release sites. *J. Physiol. (Lond.)* **494**, 231–250 (1996).
10. Chadderton, P., Margrie, T. W. & Häusser, M. Integration of quanta in cerebellar granule cells during sensory processing. *Nature* **428**, 856–860 (2004).
11. DiGregorio, D. A., Nusser, Z. & Silver, R. A. Spillover of glutamate onto synaptic AMPA receptors enhances fast transmission at a cerebellar synapse. *Neuron* **35**, 521–533 (2002).
12. Mercer, A. et al. Excitatory connections made by presynaptic cortico-cortical pyramidal cells in layer 6 of the neocortex. *Cereb. Cortex* **15**, 1485–1496 (2005).
13. Hempel, C. M., Hartman, K. H., Wang, X. J., Turrigiano, G. G. & Nelson, S. B. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* **83**, 3031–3041 (2000).
14. Iwasaki, S. & Takahashi, T. Developmental regulation of transmitter release at the calyx of Held in rat auditory brainstem. *J. Physiol. (Lond.)* **534**, 861–871 (2001).
15. Taschenberger, H., Leao, R. M., Rowland, K. C., Spirou, G. A. & von Gersdorff, H. Optimizing synaptic architecture and efficiency for high-frequency transmission. *Neuron* **36**, 1127–1143 (2002).
16. von Gersdorff, H. & Borst, J. G. Short-term plasticity at the calyx of Held. *Nature Rev. Neurosci.* **3**, 53–64 (2002).
17. Sargent, P. B., Saviane, C., Nielsen, T. A., DiGregorio, D. A. & Silver, R. A. Rapid vesicular release, quantal variability and spillover contribute to the precision and reliability of transmission at a glomerular synapse. *J. Neurosci.* **25**, 8173–8187 (2005).
18. Sola, E., Prestori, F., Rossi, P., Taglietti, V. & D'Angelo, E. Increased neurotransmitter release during long-term potentiation at mossy fibre–granule cell synapses in rat cerebellum. *J. Physiol. (Lond.)* **557**, 843–861 (2004).
19. Scheuss, V., Schneggenburger, R. & Neher, E. Separation of presynaptic and postsynaptic contributions to depression by covariance analysis of successive EPSCs at the calyx of Held synapse. *J. Neurosci.* **22**, 728–739 (2002).
20. Betz, W. J. Depression of transmitter release at the neuromuscular junction of the frog. *J. Physiol. (Lond.)* **206**, 629–644 (1970).
21. Markram, H., Wang, Y. & Tsodyks, M. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl Acad. Sci. USA* **95**, 5323–5328 (1998).
22. Xu-Friedman, M. A. & Regehr, W. G. Ultrastructural contributions to desensitization at cerebellar mossy fiber to granule cell synapses. *J. Neurosci.* **23**, 2182–2192 (2003).
23. Schneggenburger, R., Meyer, A. C. & Neher, E. Released fraction and total size of a pool of immediately available transmitter quanta at a calyx synapse. *Neuron* **23**, 399–409 (1999).
24. Palay, S. L. & Chan-Palay, V. *Cerebellar Cortex: Cortex and Organization* (Springer, New York, 1974).
25. Hallermann, S., Pawlu, C., Jonas, P. & Heckmann, M. A large pool of releasable

- vesicles in a cortical glutamatergic synapse. *Proc. Natl Acad. Sci. USA* **100**, 8975–8980 (2003).
26. Royle, S. J. & Lagnado, L. Endocytosis at the synaptic terminal. *J. Physiol. (Lond.)* **553**, 345–355 (2003).
 27. Hamori, J. & Somogyi, J. Differentiation of cerebellar mossy fiber synapses in the rat: a quantitative electron microscope study. *J. Comp. Neurol.* **220**, 365–377 (1983).
 28. Xu-Friedman, M. A. & Regehr, W. G. Structural contributions to short-term synaptic plasticity. *Physiol. Rev.* **84**, 69–85 (2004).
 29. Zenisek, D., Steyer, J. A. & Almers, W. Transport, capture and exocytosis of single synaptic vesicles at active zones. *Nature* **406**, 849–854 (2000).
 30. Sakaba, T. & Neher, E. Calmodulin mediates rapid recruitment of fast-releasing synaptic vesicles at a calyx-type synapse. *Neuron* **32**, 1119–1131 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. DiGregorio and I. Kleppe for discussions; and D. Attwell, D. DiGregorio, I. Kleppe, T. Nielsen, A. Roth and V. Steuber for comments on the manuscript. This work was supported by The Wellcome Trust and the European Community. R.A.S. is in receipt of a Wellcome Trust Senior Fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A.S. (a.silver@ucl.ac.uk).

LETTERS

NMDA receptors mediate calcium accumulation in myelin during chemical ischaemia

I. Micu¹, Q. Jiang¹, E. Coderre¹, A. Ridsdale¹, L. Zhang¹, J. Woulfe¹, X. Yin², B. D. Trapp², J. E. McRory³, R. Rehak³, G. W. Zamponi³, W. Wang¹ & P. K. Stys¹

Central nervous system myelin is a specialized structure produced by oligodendrocytes that ensheaths axons, allowing rapid and efficient saltatory conduction of action potentials¹. Many disorders promote damage to and eventual loss of the myelin sheath, which often results in significant neurological morbidity. However, little is known about the fundamental mechanisms that initiate myelin damage, with the assumption being that its fate follows that of the parent oligodendrocyte. Here we show that NMDA (*N*-methyl-*D*-aspartate) glutamate receptors mediate Ca^{2+} accumulation in central myelin in response to chemical ischaemia *in vitro*. Using two-photon microscopy, we imaged fluorescence of the Ca^{2+} indicator X-rhod-1 loaded into oligodendrocytes and the cytoplasmic compartment of the myelin sheath in adult rat optic nerves. The AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid)/kainate receptor antagonist NBQX² completely blocked the ischaemic Ca^{2+} increase in oligodendroglial cell bodies, but only modestly reduced the Ca^{2+} increase in myelin. In contrast, the Ca^{2+} increase in myelin was abolished by broad-spectrum NMDA receptor antagonists (MK-801, 7-chlorokynurenic acid, D-AP5^{3,4}), but not by more selective blockers of NR2A and NR2B subunit-containing receptors (NVP-AAM077⁵ and ifenprodil^{2,4}). *In vitro* ischaemia causes ultrastructural damage to both axon cylinders and myelin⁶. NMDA receptor antagonism greatly reduced the damage to myelin. NR1, NR2 and NR3 subunits were detected in myelin by immunohistochemistry and immunoprecipitation, indicating that all necessary subunits are present for the formation of functional NMDA receptors. Our data show that the mature myelin sheath can respond independently to injurious stimuli. Given that axons are known to release glutamate⁷⁻⁹, our finding that the Ca^{2+} increase was mediated in large part by activation of myelinic NMDA receptors suggests a new mechanism of axo-myelinic signalling. Such a mechanism may represent a potentially important therapeutic target in disorders in which demyelination is a prominent feature, such as multiple sclerosis, neurotrauma, infections (for example, HIV encephalomyelopathy) and aspects of ischaemic brain injury.

We measured $[\text{Ca}^{2+}]$ changes from the cytosolic compartment of compact myelin (the space of compact myelin between the cytoplasmic leaflets of the spiralled myelin membranes, the 'major dense line') and oligodendroglial cell bodies in live adult rat optic nerves (Fig. 1). Nerves were rendered chemically ischaemic by using the mitochondrial inhibitor NaN_3 and omitting glucose. Ca^{2+} -dependent fluorescence from myelin regions ($F_{\text{Ca,my}}$) began to increase within minutes of the onset of ischaemia and rose to $50 \pm 22\%$ above baseline after 30 min (mean $\pm$ s.d., $n = 94$ myelin regions from individual axons). Ca^{2+} -dependent fluorescence from oligodendroglial cytoplasm ($F_{\text{Ca,ol}}$) rose with a similar time course,

increasing by $43 \pm 27\%$ ($n = 53$) above control at 30 min. Removal of bath Ca^{2+} abolished both the myelinic and oligodendroglial ischaemic Ca^{2+} increase (Fig. 1f), indicating that in both compartments, Ca^{2+} increases are mainly a result of influx from the extracellular space.

We then explored potential routes of Ca^{2+} entry into both oligodendrocytes and myelin. Kynurenic acid (1 mM), a broad-spectrum antagonist of ionotropic glutamate receptors, completely blocked the increase in $F_{\text{Ca,ol}}$ ($-8 \pm 8\%$ of baseline at 30 min ischaemia, $n = 10$, $P = 1.6 \times 10^{-6}$ versus drug-free ischaemia) and $F_{\text{Ca,my}}$ ($-1 \pm 16\%$, $n = 29$, $P = 2.0 \times 10^{-6}$). The more selective AMPA/kainate receptor antagonist NBQX (30 μM) completely blocked the rise in $F_{\text{Ca,ol}}$ ($-5 \pm 11\%$ of control at 30 min, $n = 33$, $P = 1.6 \times 10^{-6}$) but had only a modest effect on myelinic Ca^{2+} rise ($40 \pm 18\%$, $n = 57$ in NBQX versus $50 \pm 22\%$ in drug-free ischaemia, $P = 0.003$; Fig. 2). This finding has two important implications. First, the fact that an AMPA/kainate receptor antagonist is able to dissociate the ischaemic Ca^{2+} increase in the two compartments indicates that the myelinic Ca^{2+} increase is not merely a passive phenomenon attributable to diffusion of Ca^{2+} from the parent cell body, but is instead mediated by a distinct mechanism. Second, the pharmacological profile suggests an unexpected route of Ca^{2+} influx into myelin. Although AMPA and kainate receptors are known to be present and functional in white matter glia, NMDA receptors are generally thought to be absent¹⁰; yet the above results suggest that NMDA receptors mediate ischaemic Ca^{2+} rise in the myelin sheath.

To explore this possibility more directly, we treated ischaemic optic nerves with selective NMDA receptor antagonists. The non-competitive channel blocker MK-801 (10 μM), and competitive glutamate (D-AP5, 100 μM) and glycine (7-chlorokynurenic acid, 200–500 μM) site antagonists³, completely blocked the increase in myelinic Ca^{2+} ($-1 \pm 19\%$, $n = 50$ for MK-801; $-16 \pm 17\%$, $n = 103$ for D-AP5; $-20 \pm 14\%$, $n = 45$ for 7-chlorokynurenic acid at 30 min, versus $50 \pm 22\%$ without drug; $P = 2.1 \times 10^{-6}$ for all groups; Fig. 2c). All three agents also reduced the ischaemic Ca^{2+} increase in oligodendrocyte cell bodies ($20 \pm 35\%$, $n = 20$, $P = 0.0002$; $-20 \pm 20\%$, $n = 49$, $P = 1.8 \times 10^{-6}$; $-22 \pm 23\%$, $n = 23$, $P = 1.8 \times 10^{-6}$, respectively, versus $43 \pm 27\%$ without drug), suggesting that NMDA receptors may also directly admit Ca^{2+} into these cells. This is consistent with the observation of NMDA-induced currents in rat spinal cord oligodendrocytes¹¹.

Functional NMDA receptors are heteromers composed of NR1 and NR2 subunits, which express the glycine and glutamate recognition sites, respectively^{2,4}. Discovered more recently, the NR3 subunit may serve a modulatory role¹². A number of splice variants have been identified for each subunit, resulting in a potentially very diverse family of receptors. More specific blockers of NR2A- and

¹Ottawa Health Research Institute, Division of Neuroscience and Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario K1Y 4K9, Canada.

²Department of Neurosciences, Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA. ³Department of Physiology and Biophysics, Hotchkiss Brain Institute, University of Calgary, Alberta T2N 4N1, Canada.

NR2B-containing receptors (NVP-AAM077⁵, tested at 0.4–5 μ M; ifenprodil⁴, 10 μ M) did not significantly reduce the ischaemic increase in $F_{Ca,my}$ ($43 \pm 14\%$, $n = 22$ for 0.4 μ M NVP; $63 \pm 34\%$, $n = 28$ for ifenprodil, versus $50 \pm 22\%$ without antagonist) (Fig. 2c). Taken together, our pharmacological data suggest that NMDA receptors (possibly those not composed of NR2A or NR2B subunits) mediate the bulk of pathological Ca^{2+} accumulation into the myelin sheath of adult optic axons. This is in contrast to oligodendrocytes, in which Ca^{2+} increase and cell death are mediated mainly, but possibly not exclusively (see above), by glutamate receptors of the AMPA/kainate class as previously reported¹³ (but see ref. 14).

We next used immunohistochemistry to localize NMDA receptor subunits (Fig. 3). NR1 subunits were detected just outside axon cylinders. Because virtually all adult rat optic axons are myelinated, regions immediately adjacent to axon cylinders represent myelin. NR2 and NR3 showed a similar distribution. Staining for all three subunits often appeared in discrete high densities in myelin of larger

axons. Immunoelectron microscopy detected labelling of myelin internodes with antibodies against NR1, NR2 and NR3 (Fig. 3e–g), with gold particles conspicuous in the inner and outer margins of myelin. Furthermore, NR1 was detected on the rough endoplasmic reticulum and Golgi membranes located in the perinuclear cytoplasm of oligodendrocytes (Fig. 3h), supporting the idea of NR1 synthesis by the myelin-forming cell.

We extracted the myelin fraction from whole optic nerve on a sucrose gradient (see Supplementary Methods). NR1, NR2 and NR3

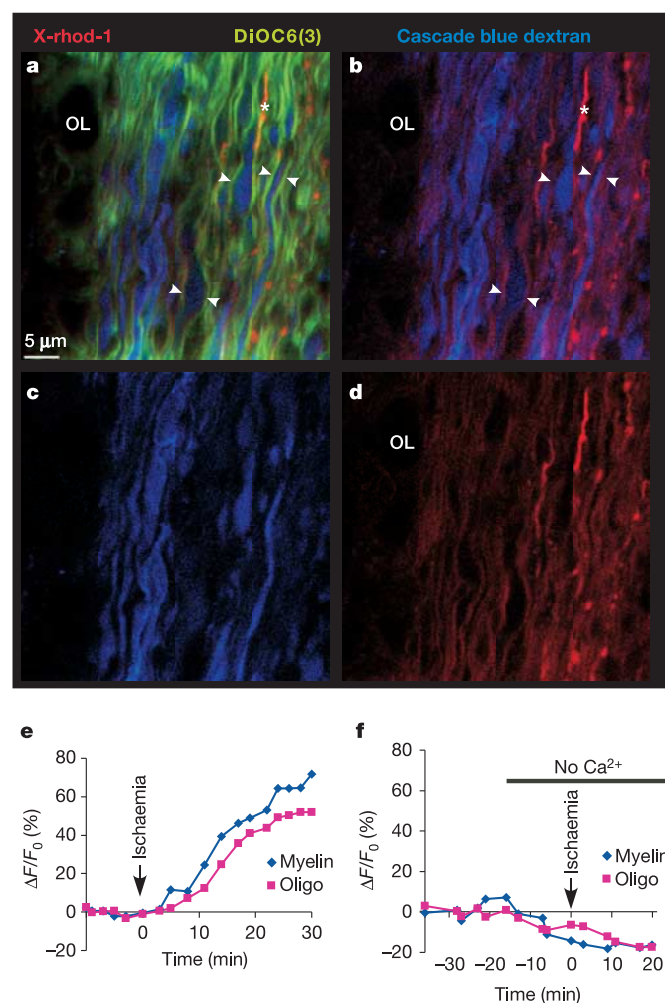


Figure 1 | Ca^{2+} imaging in optic nerve oligodendrocytes and myelin. **a–d**, The same image is shown, but with various colour channels suppressed. Cascade blue dextran outlines axon cylinders, DiOC6(3) strongly stains myelin (arrowheads in **a**), and X-rhod-1 partitions into myelin (arrowheads in **b**) and oligodendrocytes (OL). Glial processes unstained by DiOC6(3) (asterisk in **a** and **b**) were excluded. **e**, Representative time course of X-rhod-1 fluorescence change in oligodendrocytes and myelin. Fluorescence began to increase in both compartments within minutes of ischaemia, rising to ~50–60% above control levels after about 30 min. **f**, This increase was prevented in Ca^{2+} -free aCSF (with 1 mM EGTA added to chelate residual Ca^{2+}), indicating that Ca^{2+} increase in oligodendrocytes and myelin depends on influx from the extracellular space.

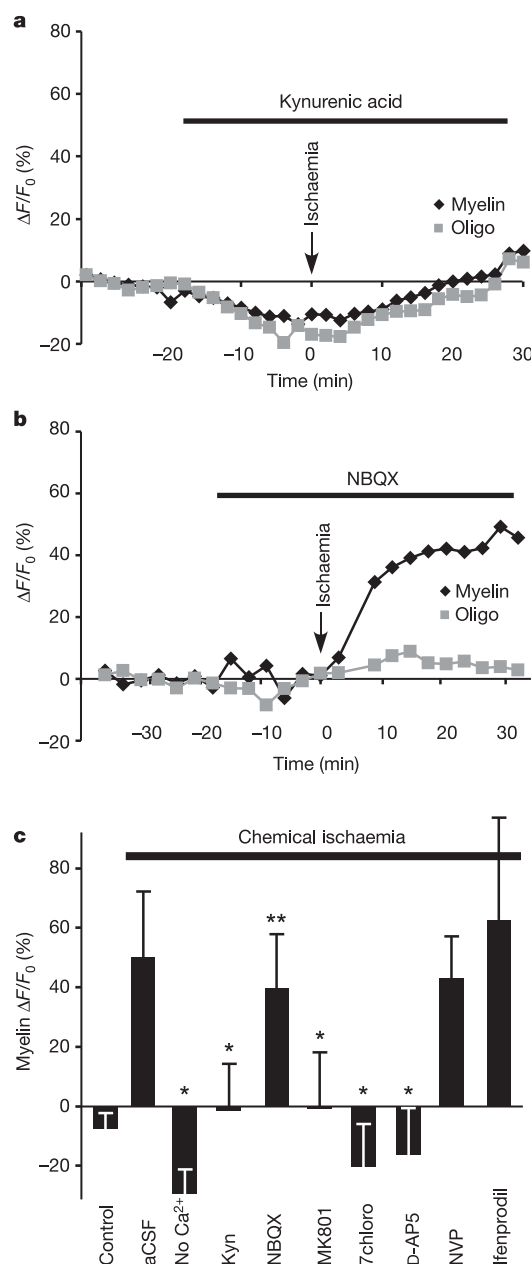


Figure 2 | Effects of glutamate receptor antagonists on ischaemic Ca^{2+} increase. **a**, Kynurenic acid greatly reduced the ischaemic Ca^{2+} increase in both oligodendrocytes and myelin (compare to Fig. 1e). **b**, NBQX blocked the Ca^{2+} increase in oligodendrocytes much more potently than in myelin. **c**, Bar graph summarizing X-rhod-1 fluorescence change (mean $\pm$ s.d.) at 30 min of ischaemia compared to pre-ischaemic/pre-drug baseline. Together, these data indicate that NMDA receptors are largely responsible for myelinic Ca^{2+} accumulation. * $P < 10^{-5}$; ** $P = 0.003$ versus drug-free ischaemia (aCSF) ($\pm$ DMSO as applicable). For each treatment group, 29–104 myelin regions of interest from at least three different experiments were analysed.

subunits could not be detected in the unpurified fraction (not shown). After immunoprecipitation, however, we were able to detect NR1-, NR2- and NR3-positive bands at the expected molecular masses by immunoblotting. Moreover, NR2 and NR3 subunits were detected in lysate precipitated by NR1 antibodies from the myelin fraction (Fig. 3i). We note that the NR2 blot revealed a band of approximately 140 kDa, which, given the known cross-reactivity of the antibody used (see Supplementary Methods), could be consistent with NR2C and/or NR2D subunits. Overall, our biochemical analysis supports the hypothesis that myelin expresses functional NMDA receptor complexes.

The expression of functional NMDA receptors in the mature myelin sheath and their ability to mediate ischaemic Ca^{2+} accumulation suggests that these receptors may have an important role in the degeneration of myelin, and may be an early trigger for eventual demyelination. Examination of ischaemic optic nerve axons by electron microscopy showed marked destruction of axoplasmic constituents, with disappearance of microtubules and neurofilament profiles⁶ (Fig. 4). In addition, the myelin sheath began to split focally (Fig. 4a, arrowheads), with loosening and separation of the lamellae. NMDA receptor inhibition significantly reduced the pathological changes in the myelin sheath; in contrast, there was little evidence that axon cylinders *per se* were protected, indicating that these elements are damaged by another mechanism (Fig. 4b). This finding is consistent with our observation that NMDA receptor block failed to blunt axoplasmic (as opposed to myelinic) Ca^{2+} increase during ischaemia (Supplementary Fig. 3).

Myelin is a vital structure that has been observed histopathologically to degenerate in a broad range of CNS disorders¹⁵. These demyelinated central axons conduct aberrantly or not at all, and account for significant clinical morbidity. During development,

reciprocal axo–glial signalling has a key role in the initiation and completion of myelination¹⁶. In contrast, little is known about the signalling that may take place between axons and the mature myelin sheath. Numerous studies have demonstrated the presence of many enzymes¹⁷ (including Na^+ , K^+ -ATPase¹⁸), connexins¹⁹ and AMPA/kainate receptors^{20,21} in mature myelin, suggesting that this structure may be more dynamic and metabolically active than previously thought, actively participating in signal transduction and translocation of ions and small molecules. Here we show that mature myelin is compromised by ischaemia and that Ca^{2+} ions accumulate in its cytosolic domain in an NMDA receptor-dependent fashion. Given the presence of Ca^{2+} -sensitive enzymes such as calpain-1 and phospholipase C in central myelin^{17,22}, it is possible that such Ca^{2+} accumulation then promotes degradation of a number of key structural myelin components.

NMDA receptors must bind both glutamate and glycine as obligatory co-agonists, and activation additionally requires depolarization to remove Mg^{2+} -dependent block^{2,23}. Both neurotransmitters are taken back up into cells by electrogenic, Na^+ -coupled transporters, which can operate in the reverse, transmitter efflux mode under conditions of increased intracellular $[\text{Na}^+]$ and membrane depolarization^{24,25}. Injured axons accumulate Na^+ , lose K^+ and depolarize (see ref. 26 for a review), thus promoting the release of glutamate via this mechanism⁸. CNS myelinated axons have been shown to express glycine transporters as well²⁷; therefore they could release this amino acid under pathological conditions in a manner analogous to glutamate. Together with ischaemic depolarization of cellular membranes, this would complete the requirements for sustained and deleterious activation of myelinic NMDA receptors.

NMDA receptors have fundamental roles in the normal physiology of the CNS²⁸ and in many diseases^{4,29}. These receptors have not been

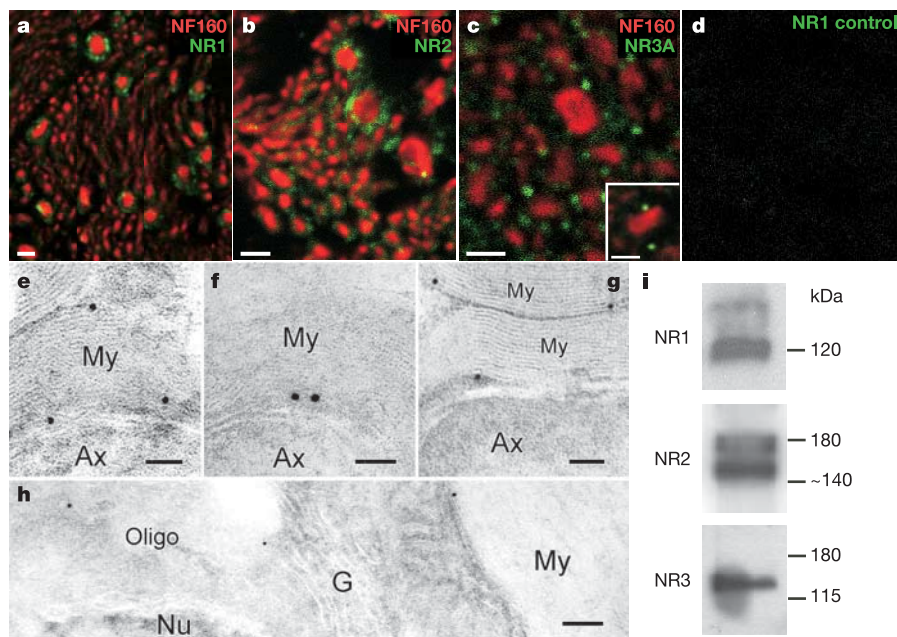


Figure 3 | Immunodetection of NMDA receptor subunits in myelin.

a, Nerves were immunostained for neurofilament (NF)160 and NR1 subunits, and imaged by confocal microscopy. Most axons in adult nerves are myelinated, and therefore regions immediately outside axon cylinders represent myelin sheaths. Faint, homogeneous NR1 signal was seen just outside NF160-positive axon cylinders, and was most apparent in larger axons. Myelin of larger-diameter axons showed occasional punctate regions of more intense NR1 label. **b**, **c**, NR2 and NR3 subunits showed a similar distribution. Inset in **c** shows an example of punctate NR3A staining at the inner and outer myelin rim of a larger axon. **d**, Control with primary antibodies omitted. Controls for NR2 and NR3 were similar.

e–g, Immunogold staining of optic nerve for NR1 (**e**), NR2 (**f**) and NR3 (**g**) showed specific myelin labelling (My), with enrichment of gold particles at the inner and outer margins of the sheath (Ax, axon). **h**, NR1 was also detected on the rough endoplasmic reticulum and Golgi membranes (G) in oligodendrocyte perinuclear cytoplasm (Nu, nucleus). **i**, A pan-NR1 monoclonal was used to immunoprecipitate receptor complexes from myelin fractions in all three blots, which were then probed with anti-NR1, anti-NR2 or anti-NR3 antisera. All three subunits were detected, and both NR2 and NR3 are complexed with NR1. Scale bars, 2 μm (**a–c**), 50 nm (**e–g**), 100 nm (**h**).

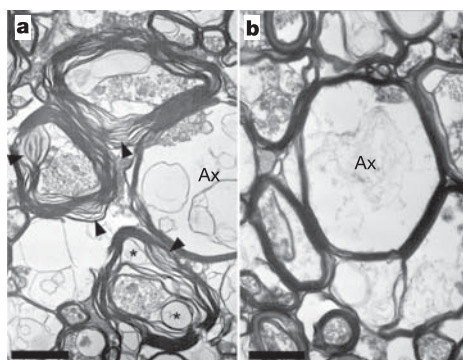


Figure 4 | NMDA receptor inhibition protects myelin against ischaemic injury. **a**, Vehicle-treated nerves showed marked damage to myelin after 60 min of *in vitro* chemical ischaemia, with separation of the compact lamellae (arrowheads), swellings and vacuolations (asterisks). Axon cylinders (Ax) were also severely damaged and swollen, with the disappearance of neurofilament and microtubule profiles. **b**, The NMDA receptor antagonist 7-chlorokynurenic acid (500 μ M) protected myelin. Occasional focal loosening of the compact sheath was observed, but was much less pronounced compared to controls (2.0 ± 1.0 lamellar separations per $65 \mu\text{m}^2$ in 7-chlorokynurenic acid versus 4.0 ± 1.4 in vehicle alone, $P = 0.0016$). In contrast, damage to axon cylinders (Ax) remained severe. Scale bars, 1 μ m.

considered important in the physiology or pathobiology of white matter, but it is likely that their presence in myelin suggests a normal physiological role. Action potential traffic releases glutamate from axons^{7,9}, and glycine could also be released in an activity-dependent manner. If myelinic NMDA receptors are composed of NR2C and/or NR2D subunits, as our data suggest, they would be more sensitive to glutamate and glycine^{2,4}. This would assist signalling in a system (myelinated axons) that is, in contrast to synapses, presumably less able to release large local amounts of transmitter. The activity-dependent release of axonal glutamate, in quantities reflecting an integral of ongoing electrical traffic in the sourcing fibre, might activate myelinic NMDA receptors to support axo–glial communication. This would represent an elegant solution for allowing axons, which may be conducting different volumes of impulse traffic, to selectively signal their myelin sheaths about their levels of activity. Such a signal could in turn be transduced to the parent oligodendrocyte for purposes of myelin production commensurate with the degree of electrical activity of individual fibres. Under pathophysiological conditions, excessive stimulation of myelinic NMDA receptors may cause direct Ca^{2+} -dependent damage to this structure. For instance, the NMDA receptor antagonist memantine²⁹ was reported to reduce neurological disability in a rat model of experimental autoimmune encephalomyelitis³⁰. Although the mechanisms are not clear, it is possible that this agent reduced myelin damage by antagonizing the effects of glutamate released in the inflammatory lesions. A more in-depth understanding of such an NMDA-receptor-dependent ‘axo–myelinic’ communication could guide the design of more effective treatments for disorders in which demyelination of central white matter tracts is a prominent and clinically devastating phenomenon.

METHODS

Please refer to Supplementary Information for additional details.

Ca^{2+} imaging. Rat optic nerves from adult Long Evans rats were loaded with the Ca^{2+} indicator X-rhod-1 AM to report the free $[\text{Ca}^{2+}]$ in myelin and oligodendrocytes, and with DiOC6(3) to outline myelin. Dextran-conjugated Ca^{2+} -insensitive cascade blue was selectively loaded into axons. Two-photon excited fluorescence images were collected every 2 min using a custom-modified Nikon D-Eclipse C1 laser scanning microscope. Emitted fluorescence was collected using appropriate filters. Image data were imported into ImageTrak software (written by P.K.S.; <http://www.ohri.ca/stys/imagetrak>) for analysis. All

quantitative fluorescence changes are reported after a standard 30-min exposure to ischaemia. In-place immunohistochemistry of live nerves was performed to ascertain the localization of X-rhod-1. Fluorescence-lifetime measurements confirmed that fluorescence intensity of X-rhod-1 reported $[\text{Ca}^{2+}]$ changes, and Mn^{2+} quench studies showed that emission originated from the cytosolic compartment of myelin (see Supplementary Methods).

Immunohistochemistry and immunoelectron microscopy. For light microscopy, deeply anaesthetized adult rats were perfused with saline then 4% paraformaldehyde in 0.1 M phosphate buffer. Optic nerves were post-fixed, placed in 20% sucrose overnight, then sectioned at a thickness of 30 μ m in a cryostat. All washes were done in TBS/1% Triton X-100, and sections were blocked with 10% normal donkey serum, 3% dry milk for 1 h at 20 °C. Samples were incubated overnight in primary antibodies, then secondary antibodies were applied. Sections were imaged on a Nikon D-Eclipse C1 confocal microscope. Immunoelectron microscopy was performed using standard techniques as described in Supplementary Methods.

Immunoblots and immunoprecipitation. Myelin fractions were incubated on ice for 1 h, centrifuged for 15 min (4 °C) at 15,000g, and the supernatant was collected for overnight dialysis against the binding solution (300 mM NaCl, 50 mM Tris pH 7.5 and 0.1% Triton X-100). Total protein (250 μ g in 50 μ l) was immunoprecipitated with a pan-NR1 monoclonal antibody. The complex was centrifuged, resuspended and loaded onto a 6% acrylamide SDS-PAGE gel. Samples were transferred to a PVDF membrane and western blot analysis was performed using a different NR1 polyclonal antibody, or NR2 or NR3 antisera. Appropriate negative controls were performed by omitting the primary (precipitating) antibody or probing with secondary antibody only (data not shown).

Received 20 August; accepted 29 November 2005.

Published online 21 December 2005.

- Waxman, S. G., Kocsis, J. D. & Stys, P. K. *The Axon: Structure, Function and Pathophysiology* (Oxford Univ. Press, Oxford, 1995).
- Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. F. The glutamate receptor ion channels. *Pharmacol. Rev.* **51**, 7–61 (1999).
- Waxman, E. A. & Lynch, D. R. N-methyl-D-aspartate receptor subtypes: multiple roles in excitotoxicity and neurological disease. *Neuroscientist* **11**, 37–49 (2005).
- Parsons, C. G., Danysh, W. & Quack, G. Glutamate in CNS disorders as a target for drug development: an update. *Drug News Perspect.* **11**, 523–569 (1998).
- Auberson, Y. P. *et al.* 5-phosphonomethylquinolinediones as competitive NMDA receptor antagonists with a preference for the human 1A/2A, rather than 1A/2B receptor composition. *Bioorg. Med. Chem. Lett.* **12**, 1099–1102 (2002).
- Waxman, S. G., Black, J. A., Stys, P. K. & Ransom, B. R. Ultrastructural concomitants of anoxic injury and early post-anoxic recovery in rat optic nerve. *Brain Res.* **574**, 105–119 (1992).
- Lin, S. C. *et al.* Climbing fiber innervation of NG2-expressing glia in the mammalian cerebellum. *Neuron* **46**, 773–785 (2005).
- Li, S., Mealing, G. A., Morley, P. & Stys, P. K. Novel injury mechanism in anoxia and trauma of spinal cord white matter: glutamate release via reverse Na^{+} -dependent glutamate transport. *J. Neurosci.* **19**, RC16 (1999).
- Chiu, S. Y. & Kriegler, S. Neurotransmitter-mediated signalling between axons and glial cells. *Glia* **11**, 191–200 (1994).
- Matute, C., Alberdi, E., Ibarretxe, G. & Sanchez-Gomez, M. V. Excitotoxicity in glial cells. *Eur. J. Pharmacol.* **447**, 239–246 (2002).
- Ziak, D., Chvatal, A. & Sykova, E. Glutamate-, kainate- and NMDA-evoked membrane currents in identified glial cells in rat spinal cord slice. *Physiol. Res.* **47**, 365–375 (1998).
- Das, S. *et al.* Increased NMDA current and spine density in mice lacking the NMDA receptor subunit NR3A. *Nature* **393**, 377–381 (1998).
- McDonald, J. W., Althomsons, S. P., Hyrc, K. L., Choi, D. W. & Goldberg, M. P. Oligodendrocytes from forebrain are highly vulnerable to AMPA/kainate receptor-mediated excitotoxicity. *Nature Med.* **4**, 291–297 (1998).
- Rosenberg, P. A. *et al.* Mature myelin basic protein-expressing oligodendrocytes are insensitive to kainate toxicity. *J. Neurosci.* **71**, 237–245 (2003).
- Lazzarini, R. A. *et al.* *Myelin Biology and Disorders* (Elsevier, Amsterdam, 2004).
- Michailov, G. V. *et al.* Axonal neuregulin-1 regulates myelin sheath thickness. *Science* **304**, 700–703 (2004).
- Quarles, R. H., Macklin, W. B. & Morell, P. in *Basic Neurochemistry* (eds Siegel, G. J., Albers, R. W., Brady, S. T. & Price, D. L.) 51–72 (Elsevier, San Diego, in the press).
- Mrsulja, B. J., Zalewski, A. A. & Copping, G. Ultracytochemical localization of ouabain-sensitive K^{+} -dependent, p-nitrophenyl phosphatase in myelin. *Brain Res.* **343**, 154–158 (1985).
- Kamasawa, N. *et al.* Connexin-47 and connexin-32 in gap junctions of oligodendrocyte somata, myelin sheaths, paranodal loops and

- Schmidt-Lanterman incisures: implications for ionic homeostasis and potassium siphoning. *Neuroscience* **136**, 65–86 (2005).
20. Brand-Schieber, E. & Werner, P. AMPA/kainate receptors in mouse spinal cord: cell-specific display of receptor subunits by oligodendrocytes and astrocytes and at the nodes of Ranvier. *Glia* **42**, 12–24 (2003).
21. Li, S. & Stys, P. K. Mechanisms of ionotropic glutamate receptor-mediated excitotoxicity in isolated spinal cord white matter. *J. Neurosci.* **20**, 1190–1198 (2000).
22. Hinman, J. D., Duce, J. A., Siman, R. A., Hollander, W. & Abraham, C. R. Activation of calpain-1 in myelin and microglia in the white matter of the aged rhesus monkey. *J. Neurochem.* **89**, 430–441 (2004).
23. Mayer, M. L., Westbrook, G. L. & Guthrie, P. B. Voltage-dependent block by Mg^{2+} of NMDA responses in spinal cord neurones. *Nature* **309**, 261–263 (1984).
24. Aragon, C. & Lopez-Corcuera, B. Structure, function and regulation of glycine neurotransmitters. *Eur. J. Pharmacol.* **479**, 249–262 (2003).
25. Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65**, 1–105 (2001).
26. Stys, P. K. White matter injury mechanisms. *Curr. Mol. Med.* **4**, 113–130 (2004).
27. Spike, R. C., Watt, C., Zafra, F. & Todd, A. J. An ultrastructural study of the glycine transporter GLYT2 and its association with glycine in the superficial laminae of the rat spinal dorsal horn. *Neuroscience* **77**, 543–551 (1997).
28. Bear, M. F. & Malenka, R. C. Synaptic plasticity: LTP and LTD. *Curr. Opin. Neurobiol.* **4**, 389–399 (1994).
29. Lipton, S. A. & Rosenberg, P. A. Excitatory amino acids as a final common pathway for neurologic disorders. *N. Engl. J. Med.* **330**, 613–622 (1994).
30. Wallstrom, E. *et al.* Memantine abrogates neurological deficits, but not CNS inflammation, in Lewis rat experimental autoimmune encephalomyelitis. *J. Neurol. Sci.* **137**, 89–96 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the NINDS, CIHR, Heart and Stroke Foundation of Ontario Center for Stroke Recovery, an HSFO Career Investigator Award, the Canadian Institute for Photonic Innovations, and the generosity of private donors to P.K.S. G.W.Z. is a Canada Research Chair and an AHFMR Senior Scholar. This work was additionally supported by the CIHR (G.W.Z., J.E.M.) and NINDS (B.D.T.). We thank Y. Auberson for the gift of NVP-AAM077; J. A. Wang and C. E. Morris for providing HEK293 cells for fluorescence-lifetime measurements; and M. Nikolaeva for assistance with fluo-4 Ca^{2+} imaging.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.K.S. (pstys@ohri.ca).

Purification and unique properties of mammary epithelial stem cells

John Stingl^{1,3}, Peter Eirew¹, Ian Ricketson¹, Mark Shackleton⁴, François Vaillant⁴, David Choi¹, Haiyan I. Li² & Connie J. Eaves^{1,5}

Elucidation of the cellular and molecular mechanisms that maintain mammary epithelial tissue integrity is of broad interest and paramount to the design of more effective treatments for breast cancer¹. Evidence from both *in vitro* and *in vivo* experiments suggests that mammary cell differentiation is a hierarchical process originating in an uncommitted stem cell with self-renewal potential^{2–4}. However, analysis of the properties and regulation of mammary stem cells has been limited by a lack of methods for their prospective isolation. Here we report the use of multi-parameter cell sorting and limiting dilution transplant analysis to demonstrate the purification of a rare subset of adult mouse mammary cells that are able individually to regenerate an entire mammary gland within 6 weeks *in vivo* while simultaneously executing up to ten symmetrical self-renewal divisions. These mammary stem cells are phenotypically distinct from and give rise to mammary epithelial progenitor cells that produce adherent colonies *in vitro*. The mammary stem cells are also a rapidly cycling population in the normal adult and have molecular features indicative of a basal position in the mammary epithelium.

The mammary gland develops after birth from a small number of invading cells derived from the ectoderm⁵. These cells then generate a series of branching ducts that terminate in sac-like lobules embedded in stromal tissue. The structures produced are composed of a continuous epithelium consisting of an outer basal layer of contractile myoepithelial cells and an inner layer of luminal cells. Experiments with retrovirally marked mouse mammary cells transplanted into cleared mammary fat pads have established the existence of mammary stem cells that are able to regenerate new mammary tissue *in vivo* and display self-renewal activity². Here we refer to such cells operationally as mammary repopulating units (MRUs). Mammary glands also contain different types of progenitor cells that can be detected under various conditions *in vitro*^{3,6,7}. Mammary colony-forming cells (Ma-CFCs) refer to those progenitors that produce discrete colonies of mammary cells in low-cell density adherent cultures⁷. At present, little is known about the biological properties of MRUs, or the mechanisms that regulate their behaviour and relationship to mammary cells that proliferate *in vitro*. Investigation of these key questions is essential to developing an understanding of mammary stem cell fate decisions and transformation. To begin to address these, we analysed the phenotypic and cycling properties of cells detectable *in vivo* as MRUs and compared them to cells detectable *in vitro* as Ma-CFCs.

In preliminary experiments we found that MRUs could be routinely detected in single cell suspensions prepared by enzymatic digestion of adult mouse mammary tissue and regenerated macroscopic mammary outgrowths within 6 weeks after injection into cleared mammary fat pads of weaning female mice⁸. The outgrowths produced (Fig. 1a, b)

were histologically normal (Fig. 1c) and contained both mature keratin 18-expressing luminal cells and smooth muscle actin-positive myoepithelial cells (Fig. 1d). *In vitro* assays of single cell suspensions prepared from such regenerated outgrowths demonstrated that they also contained numerous Ma-CFCs (Fig. 1e), thus demonstrating a parent–progeny relationship between MRUs and Ma-CFCs.

To investigate whether reaggregation of dissociated cells is an essential step in the formation of MRU-derived outgrowths, small numbers of cells from green fluorescent protein (GFP)⁹ and cyan fluorescent protein (CFP)¹⁰ donors were mixed together and transplanted. Most outgrowths produced from these mixed transplants were found to contain either GFP⁺ or CFP⁺ Ma-CFCs, but not both genotypes (Fig. 1e), providing strong support for the view that single cells can generate complete outgrowths. Formal proof of this concept was obtained from the results of single cell transplants using highly purified MRUs (Fig. 1f and Supplementary Fig. 1; see below for the phenotype selection strategy and MRU purity achieved). In these experiments, the purified cells were first placed at a low concentration in a multiwell plate and the contents of wells containing only one cell then selected by microscopic inspection before being individually harvested and injected into cleared fat pads. The ability of a single cell to generate an outgrowth (Fig. 1f) indicates that MRUs can display their regenerative potential in the absence of any co-injected cells.

Consistent with these findings, we observed a negative linear relationship between the transplant cell dose and the log proportion of cleared fat pads that did not contain outgrowths when the number of cells injected was varied (Fig. 1g). This relationship also validated the use of limiting dilution analysis to quantify MRUs in variously manipulated suspensions of mammary cells. The frequency of MRUs in single cell suspensions obtained from adult virgin female FVB mouse inguinal mammary glands dissociated using an 8-h enzymatic procedure was determined by limiting dilution analysis to be 1 MRU per 1,400 dissociated cells (95% confidence interval = 1 per 600 to 1 per 3,000 cells, Supplementary Table 1). The total yield of viable cells obtained by this dissociation procedure was approximately 2 million per gland. Therefore, each inguinal gland contains at least 1,400 MRUs and probably more considering the imperfect yield of single cells expected from the dissociation procedure used. Similar values were determined for the frequency and yield of MRUs in cell suspensions obtained from the mammary glands of C57Bl/6 mice (Fig. 1e), indicating the reproducibility of these values between strains. Notably, when the same inguinal mammary gland cell suspensions were assayed *in vitro*, Ma-CFCs were detected at a 20-fold higher frequency (that is, 1 Ma-CFC per 63 cells ($\pm$ s.e.m. = 1 per 45 to 1 per 108), indicating a yield of approximately 30,000 Ma-CFCs per gland) in comparison to the values obtained for

¹Terry Fox Laboratory, and ²Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver V5Z 1L3, British Columbia, Canada. ³StemCell Technologies Inc., Vancouver V5Z 1B3, British Columbia, Canada. ⁴The Walter and Eliza Hall Institute of Medical Research, WEHI Biotechnology Centre, Parkville, Victoria 3050, Australia.

⁵Departments of Medical Genetics, Pathology and Laboratory Medicine, and Medicine, University of British Columbia, Vancouver V6T 1Z3, British Columbia, Canada.

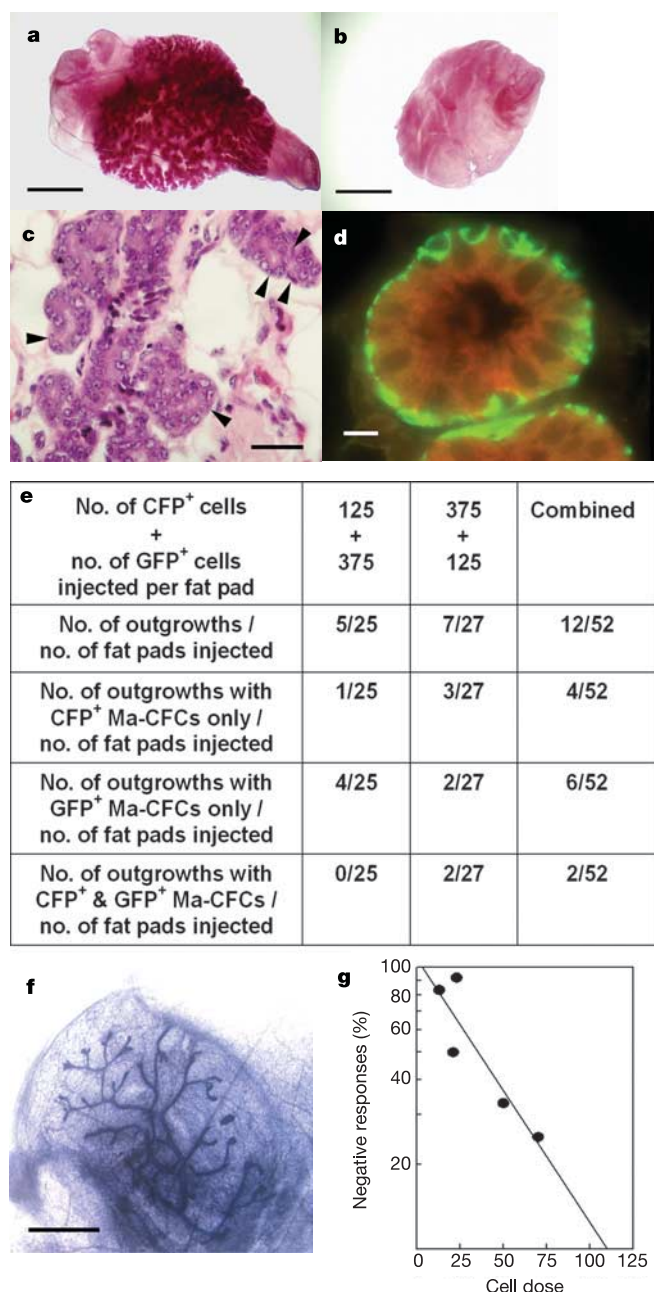


Figure 1 | The MRU assay. **a**, Mammary gland outgrowth produced in a transplanted fat pad. Scale bar, 0.25 cm. **b**, Cleared fat pad not injected with cells. Scale bar, 0.25 cm. **c**, Stratified epithelium visualized in a haematoxylin and eosin-stained gland obtained in a fat pad injected with 12 CD45⁺ Ter119⁺ CD31⁺ Sca-1^{low} CD24^{med} CD49f^{high} cells. Arrowheads indicate myoepithelial cells. Scale bar, 25 μ m. **d**, Regenerated gland showing keratin 18-positive luminal cells (red) and smooth muscle actin-positive myoepithelial cells (green). Scale bar, 10 μ m. **e**, Proportions of outgrowths produced in fat pads transplanted with a mixture of GFP⁺ and CFP⁺ cells (see Supplementary Methods) that contained GFP⁺ and/or CFP⁺ Ma-CFCs. Frequencies of MRUs in the original GFP⁺ and CFP⁺ suspensions were 1 per 1,500 (95% confidence interval = 1 per 700 to 1 per 3,000) and 1 per 2,000 (95% confidence interval = 1 per 910 to 1 per 4,500). The probability of obtaining a mixed outgrowth given these individual frequencies was 0.015 (not significantly different from the observed outcome of 2 per 52 = 0.04, $P = 0.19$). **f**, A mammary gland regenerated in a fat pad of a virgin host injected with a single CD45⁺ Ter119⁺ CD31⁺ Sca-1^{low} CD24^{med} CD49f^{high} cell (one positive fat pad of 77 injected with single cells). Scale bar, 1 mm. **g**, Negative linear relationship between the number of CD45⁺ Ter119⁺ CD31⁺ CD140a⁺ CD24^{med} CD49f^{high} mammary cells transplanted and the proportion of negative fat pads obtained. The goodness of fit of this model was confirmed ($P = 0.40$).

MRUs. These findings suggested that MRUs represent a much rarer cell type than Ma-CFCs.

To characterize MRUs further, we analysed their distribution in various subpopulations of cells from enzymatically digested mammary glands that could be isolated using the fluorescence activated cell sorter (FACS) after staining the cells with antibodies to various surface markers. Markers examined included Sca-1, a marker previously described to identify MRUs¹¹; CD24, a marker for which the absence characterizes human mammary tumour stem cells¹; and CD49f, a marker expressed by epidermal stem cells^{12,13} and human Ma-CFCs⁷. In human mammary tissue, CD24 has been localized to the apical plasma membrane of the luminal cells¹⁴, although similar studies of mouse mammary tissue suggested a more ubiquitous pattern of expression (Fig. 2a). In contrast, CD49f is most highly expressed by cells of the basal layer of skin epithelial cells¹⁵, and a basal pattern of expression was confirmed here for the mouse mammary gland (Fig. 2b). We found that when contaminating haematopoietic (CD45⁺ and Ter119⁺) cells were first removed and then the most highly CD49f⁺ (CD49f^{high}) cells were selected (Fig. 2c), either with or without simultaneous selection of the Sca-1^{low} subset (cells within the lower-right gate indicated in Fig. 2d), more than 50% of the MRUs in suspensions of FVB mouse mammary cells could be recovered at a sevenfold higher frequency (~1 per 200 cells; 4 positive fat pads out of 6 injected with 200 cells each) than the frequency of MRUs in the initial cell suspension. No MRUs were detected in the Sca-1^{high} subset of CD49f⁺ cells (0 positive fat pads out of 8 injected with 600 cells each from the upper-left gate in Fig. 2d). We did not culture the cells before staining because this induced the expression of Sca-1 on all mammary epithelial cells (Supplementary Fig. 2). Additional removal of contaminating endothelial (CD31⁺) cells and selection of the CD24⁺ subset of the CD45⁺ Ter119⁺ CD31⁺ Sca-1^{low} CD49f^{high} cells (upper population in Fig. 2e) increased the frequency of the MRUs another tenfold (to ~1 per 20 cells; 3 positive fat pads out of 4 injected with 18 of these cells each, and 4 positive fat pads out of 10 injected with 12 of these cells, and 0 positive fat pads out of 12 injected, each with 312 of the corresponding CD24^{low/med} subset, Fig. 2e). Expression of CD24 in MRUs has also been demonstrated previously¹⁶. When stromal (CD140a⁺) cells were also removed but the Sca-1 staining was omitted, approximately 80% of the MRUs were found in a subset of cells that express medium levels of CD24 and high levels of CD49f (CD45⁺ Ter119⁺ CD31⁺ CD140a⁺ CD24^{med} CD49f^{high} cells). In this subset, the frequency of MRUs was 1 per 60 cells (from FVB mice) and 1 per 90 cells (from C57Bl/6 mice) (Fig. 3a and Supplementary Table 2). Accordingly, this fraction has been labelled in Fig. 3a as 'MRU'. The few remaining MRUs were found to have a CD24^{low} CD49f^{low} phenotype. Because immunostaining (Fig. 2b) demonstrated that differentiated myoepithelial cells express both CD24 and CD49f, the CD24^{low} CD49f^{low} fraction shown in Fig. 3a was labelled as 'MYO' (anticipating that it would be enriched in its content of myoepithelial cells).

Ma-CFCs, like MRUs, were found to co-express both CD24 and CD49f (Fig. 4a). However, they differ from MRUs in that most Ma-CFCs (90%) were found in the CD24^{high} CD49f^{low} population that lacks detectable MRUs. Hence we labelled the CD24^{high} CD49f^{low} fraction shown in Fig. 3a as 'Ma-CFC'. We also examined the types of colonies produced by cells from the Ma-CFC-enriched and MRU-enriched fractions after 16 days of growth in Matrigel cultures. The colonies generated by cells from the Ma-CFC-enriched fraction had a uniformly spherical acinar structure composed of a simple cuboidal epithelium (top panels in Fig. 3b). In contrast, cells from the MRU-enriched fraction produced solid, irregular-shaped colonies and occasional large colonies with a branched ductal appearance and an irregular-shaped lumen (Fig. 3b, lower panels).

The ability to prospectively isolate distinct, highly enriched populations of MRUs and Ma-CFCs afforded an opportunity to look for differences in their gene expression profiles and to compare

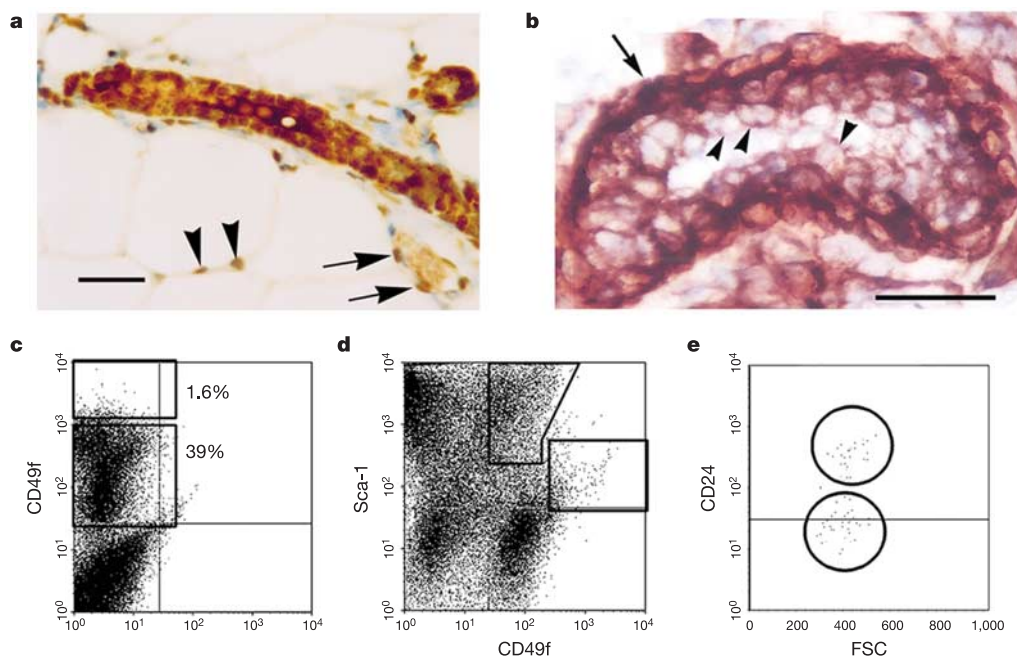


Figure 2 | Phenotypic characterization of MRUs.

a, CD24⁺ epithelial cells, adipocytes (arrowheads) and endothelial cells (arrows). Scale bar, 25 μ m. **b**, CD49f expression is higher in basal cells (arrow) than in luminal cells (arrowheads). Scale bar, 25 μ m. **c–e**, FACS dot plots showing the distribution of mammary cells according to their expression of various markers. **c**, CD49f and gates used to distinguish CD49f^{low} and CD49f^{high} cells. **d**, CD49f and Sca-1. **e**, CD24 (and forward light scattering (FSC) characteristics) within the CD45⁺Ter119⁺CD31⁺Sca-1^{low}CD49f^{high} population.

these to the MYO population that contains very few cells able to produce outgrowths in cleared fat pads or adherent colonies *in vitro*. Hybridization of amplified RNA from each of these subsets to Affymetrix mouse MOE430 genome array chips indicated that the Ma-CFC-enriched cells contain higher levels of keratin 8, 18 and 19 transcripts and a variety of casein transcripts also typical of luminal cells, in comparison to either the MRU-enriched or MYO cells (Supplementary Tables 5–7). Conversely, transcripts for keratins 5 and 14, smooth muscle actin, smooth muscle myosin, vimentin and laminin, all of which show elevated expression in basal/myoepithelial cells, were found to be present at higher levels in the MRU-enriched and MYO populations. However, significant differences in gene expression were not evident when the latter two fractions were compared. The differences in keratin 14, 18 and 19 and smooth muscle actin transcript levels in the three populations studied

were confirmed by quantitative real-time PCR analysis (Fig. 3c). Notably, transcripts for keratin 6, a putative progenitor cell marker¹⁷, were also found to be highest in the fraction enriched in Ma-CFCs (Fig. 3c).

Immunostaining of individual cells in the highly MRU-enriched fraction indicated that some expressed two markers associated with basal cells (23% positive for smooth muscle actin, $n = 620$; 27% for keratin 14, $n = 600$). Others expressed markers associated with luminal cells (18% positive for keratin 18), but none ($<0.1\%$) co-expressed markers of both lineages and approximately half did not express any of these (Fig. 3d). Immunostaining of cells isolated from the same three fractions with an antibody to keratin 6 confirmed the PCR results with 49% positive cells in the Ma-CFC-enriched fraction as compared to 0–2% positive cells in the MRU-enriched and MYO fractions (Fig. 3e).

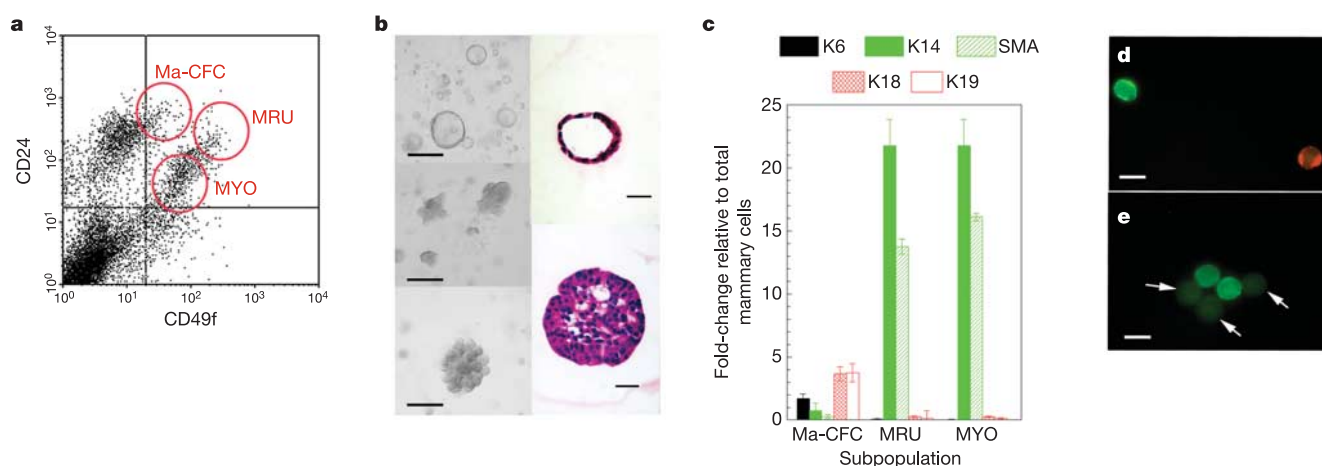


Figure 3 | Functional and molecular characterization of subsets of mammary cells. **a**, Distribution of CD45⁺Ter119⁺CD31⁺CD140a⁺ cells according to their CD24 and CD49f expression. **b**, Gross morphology (left) and haematoxylin and eosin-stained sections (right) of colonies produced in Matrigel cultures from CD24^{high}CD49f^{low} cells (Ma-CFC fraction, top panel) and from CD24^{med}CD49f^{high} cells (MRU fraction, middle and bottom panels). Scale bars, 2 mm (left panels) and 20 μ m (right panels).

c, Quantitative real-time PCR analysis of transcript levels in different subpopulations (mean $\pm$ s.e.m. of data from three independent isolates as compared to unfractionated mammary gland cells). K, keratin.

d, Immunostaining of FACS-purified CD45⁺Ter119⁺CD31⁺CD49f^{high} cells (MRU fraction) for expression of keratin 14 (green) and keratin 18 (red). Scale bar, 10 μ m. **e**, Keratin 6⁺ CD24^{high}CD49f^{low} cells (Ma-CFC fraction). Arrows indicate autofluorescence; scale bar, 10 μ m.

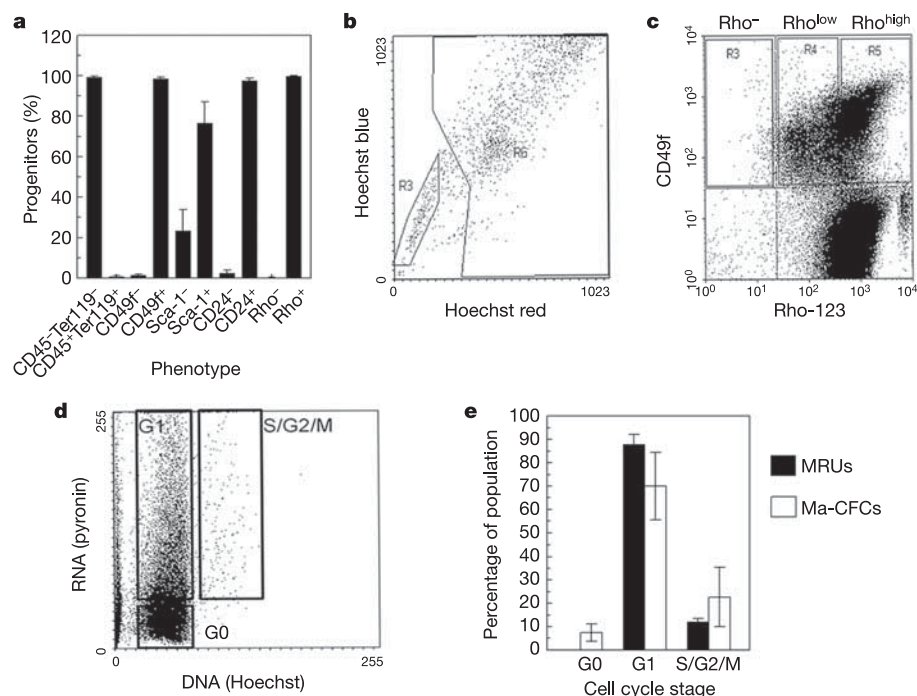


Figure 4 | Characterization of dye efflux and cycling properties of MRUs and Ma-CFCs. **a**, Percentage ($\pm$ s.e.m.) of Ma-CFCs in subpopulations indicated. **b**, Distribution of CD45⁺Ter119⁺CD49f⁺ mammary cells according to their Hoechst 33342 fluorescence (left gate, side population; right gate, non-side population). **c**, Distribution of mammary cells according to their CD49f expression and Rho efflux activity. **d**, Distribution

of CD45⁺ mammary cells in G0, G1 and S/G2/M fractions determined by Hoechst 33342/pyronin Y staining. These fractions contained $38 \pm 11\%$, $59 \pm 11\%$ and $3 \pm 1\%$ (mean $\pm$ s.e.m.), respectively, of all the CD45⁺ cells. **e**, Percentage ($\pm$ s.e.m.) of MRUs and Ma-CFCs in the G0, G1 and S/G2/M fractions.

To further characterize MRUs and Ma-CFCs, we examined their Hoechst 33342 and rhodamine-123 (Rho) efflux properties, as well as their distribution in different phases of the cell cycle. Hoechst 33342 is a substrate of Abcg2 and this allows Abcg2⁺ cells to display a unique 'side population' phenotype^{18,19}. Hoechst 33342 and Rho efflux properties are both discriminating features of quiescent murine haematopoietic stem cells²⁰ but these properties are lost when these cells are activated into cycle²¹. Primitive retinal²² and cardiac cells²³, embryonic stem cells¹⁹ and some mammary cells²⁴ have also been reported to possess a side population phenotype. We found that mammary gland preparations also contain a small fraction of side population cells ($1.9 \pm 0.5\%$ (mean $\pm$ s.e.m.) of the total, $n = 7$, Fig. 4b), but these accounted for less than 10% of all the MRUs (Supplementary Table 3). Similarly, a minority of CD49f⁺ mammary cells effluxed Rho (Fig. 4c), and this was also true for both MRUs (Supplementary Table 4) and Ma-CFCs (Fig. 4a). Analysis of Hoechst 33342/pyronin Y-stained mammary cells (Fig. 4d) showed that all of the MRUs and >95% of the Ma-CFCs were present in the G1 or S/G2/M fractions (Fig. 4e). MRUs were also broadly distributed in their forward light scattering characteristics, as expected for a cycling population (data not shown). These findings are consistent with a recent report of a cycling population of mammary epithelial cells thought to be primitive because of their ability to retain ³H-thymidine 5 weeks after a pulse exposure *in vivo*²⁵.

Self-renewal is the hallmark property of stem cells. To examine the self-renewal properties of single MRUs, 34 fat pads were transplanted with low numbers (11–42) of MRU-enriched (CD24^{med}CD49f^{high}) cells. Because outgrowths were produced in only 11 of the 34 fat pads injected, most of these could be assumed to have arisen from a single MRU. Secondary limiting dilution MRU assays were performed on cell suspensions prepared from 4 of these 11 primary outgrowths, and the results demonstrated that they contained 25, 110, 190 and 1,200 MRUs, respectively. Thus, highly purified single

MRUs could be shown to execute at least ten symmetrical self-renewal divisions.

These findings indicate a hierarchy of differentiating cells in the adult mammary gland that includes developmentally distinct and prospectively separable stem and progenitor cell types identified by quantitative *in vivo* and *in vitro* assays. The ability to isolate these cells as distinct populations sets the stage for future delineation of how mammary stem cell self-renewal and differentiation is normally regulated and perturbed during oncogenesis. The high proliferative activity of steady-state MRUs suggests an important role of apoptotic mechanisms in the normal regulation of these cells and may enhance their vulnerability to mutagenic events.

METHODS

Cell preparation. Mammary glands from 8–14-week-old virgin female FVB, C57Bl/6, or congenic GFP⁹ and CFP¹⁰ mice were digested for 8 h at 37°C in EpiCult-B with 5% fetal bovine serum (FBS), 300 U ml⁻¹ collagenase and 100 U ml⁻¹ hyaluronidase. Prolongation of the period of enzymatic digestion to 16 h resulted in a selective and eightfold overall decrease in MRU yield. After vortexing and lysis of the red blood cells in NH₄Cl, a single cell suspension was obtained by sequential dissociation of the fragments by gentle pipetting for 1–2 min in 0.25% trypsin, and then 2 min in 5 mg ml⁻¹ dispase II plus 0.1 mg ml⁻¹ DNase I (DNase; Sigma) followed by filtration through a 40- μ m mesh. All reagents were from StemCell Technologies Inc. unless otherwise specified.

Cell separation. CD45⁺/Ter119⁺, CD31⁺ and CD140a⁺ cells were removed using the EasySep biotin selection kit (StemCell Technologies) after treatment of dissociated cells with 2 μ g ml⁻¹ Fc receptor antibody (2.4G2, American Type Culture Collection) followed by 10 min with a 1:500 dilution of biotinylated StemSep murine/human chimera cocktail (StemCell Technologies), 1 μ g ml⁻¹ biotinylated anti-CD31 (clone MEC13.3, Pharmingen) and 1 μ g ml⁻¹ biotinylated anti-CD140a (clone APA5, eBioscience). In some experiments, cells were incubated for 10 min in 10% rat serum and 10 μ g ml⁻¹ 2.4G2 and then anti-CD45-allophycocyanin (APC, clone 30F11, Pharmingen), anti-CD24-R-phycoerythrin (PE, clone M1/69, Pharmingen), biotinylated anti-CD31, anti-CD49f-PE or fluorescein isothiocyanate (FITC, clone GoH3, Pharmingen), or

APC (R&D Systems) and biotinylated Sca-1 or Sca-1-PE (clone E13-161.7, Pharmingen). Apoptotic cells were excluded by elimination of annexin-V⁺ cells after staining with annexin-V-APC (Pharmingen). For assays of MRUs in sorted populations, viable cell numbers were determined by microscope counts before injecting the cells into cleared fat pads. For transplants of single CD45⁺Ter119[−]CD31[−]Sca-1^{low}CD24^{med}CD49f^{high} cells, these were first distributed at 1 cell per 3 wells in Terasaki plates (Greiner Bio-One) and visually checked before selection for injection. Cells were stained with Rho, or Hoechst 33342 and/or pyronin Y and gates set as previously described^{21,26} with the modification that EpiCult-B with 5% FBS was used during the 37°C incubation step. All sorts were performed using a FACS Vantage or FACS Aria (Becton Dickinson) and gates were set to exclude >99.9% of cells labelled with isoform-matched control antibodies conjugated with the corresponding fluorochromes.

In vitro colony assays. Most Ma-CFCs assays were performed in standard 5–6-day liquid cultures as previously described for human Ma-CFCs²⁷. To assay outgrowths for Ma-CFCs, half of the suspension obtained from each outgrowth was plated in a single dish and the colonies produced 5 days later counted using a regular microscope and genotyped using a fluorescent microscope. All colonies were exclusively GFP⁺ or CFP⁺. In selected experiments, colonies were generated in 50-μl cultures of Matrigel (Becton Dickinson) covered with 4 ml of EpiCult-B medium containing 5% FCS. After 16 days, each 50-μl Matrigel culture was fixed in 4% paraformaldehyde and then removed intact, embedded in 1% agarose, fixed again in 4% paraformaldehyde, sectioned and stained with haematoxylin and eosin.

In vivo assays in cleared mammary fat pads. Mammary glands of 21-day-old female FVB or C57Bl/6J mice were cleared of endogenous epithelium as previously described²⁸, and test cells in 5–10 μl of Hanks' balanced salt solution plus 2% FBS with 10% trypan blue (Sigma) were injected into each cleared fat pad. Three weeks after surgery, the mice were mated except in the single cell transplants and some of the mixed donor cell transplants. After a total of 6 weeks, the glands were excised and stained with carmine alum (StemCell Technologies) or haematoxylin for whole-mount analysis. In mice that were mated, only complete outgrowths containing both lobular and ductal elements were scored as a positive, although very few outgrowths containing only one of these elements were seen in several hundred transplants performed. The frequency of MRUs in the cell suspension being transplanted was calculated using Poisson statistics and the method of maximum likelihood using L-Calcul software (StemCell Technologies)²⁹. When all fat pads were either positive or negative, an estimate of the MRU frequency range was obtained using one-sided 95% Clopper Pearson limits. Goodness of fit to a single-hit Poisson model was tested with standard Chi-squared statistics.

See Supplementary Methods for details on immunocytochemistry, microarray and quantitative real-time PCR analyses.

Received 12 October; accepted 22 November 2005.

Published online 4 January 2006.

- Al Hajj, M. & Clarke, M. F. Self-renewal and solid tumour stem cells. *Oncogene* **23**, 7274–7282 (2004).
- Kordon, E. C. & Smith, G. H. An entire functional mammary gland may comprise the progeny from a single cell. *Development* **125**, 1921–1930 (1998).
- Dontu, G. *et al.* In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* **17**, 1253–1270 (2003).
- Stingl, J., Raouf, A., Emsman, J. T. & Eaves, C. J. Epithelial progenitors in the normal human mammary gland. *J. Mamm. Gland Biol. Neoplasia* **10**, 49–59 (2005).
- Sakakura, T. in *The Mammary Gland: Development, Regulation, and Function* (eds Neville, M. C. & Daniel, C. W.) 37–66 (Plenum, New York, 1987).
- Smalley, M. J., Titley, J. & O'Hare, M. J. Clonal characterization of mouse mammary luminal epithelial and myoepithelial cells separated by fluorescence-activated cell sorting. *In Vitro Cell. Dev. Biol. Anim.* **34**, 711–721 (1998).
- Stingl, J., Eaves, C. J., Zandieh, I. & Emsman, J. T. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res. Treat.* **67**, 93–109 (2001).
- Smith, G. H. Experimental mammary epithelial morphogenesis in an *in vivo* model: evidence for distinct cellular progenitors of the ductal and lobular phenotype. *Breast Cancer Res. Treat.* **39**, 21–31 (1996).
- Wagers, A. J., Sherwood, R. I., Christensen, J. L. & Weissman, I. L. Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science* **297**, 2256–2259 (2002).
- Hadjantonakis, A. K., Macmaster, S. & Nagy, A. Embryonic stem cells and mice expressing different GFP variants for multiple non-invasive reporter usage within a single animal. *BMC Biotechnol.* **2**, 11 (2002).
- Wiem, B. E. *et al.* Sca-1^{pos} cells in the mouse mammary gland represent an enriched progenitor cell population. *Dev. Biol.* **245**, 42–56 (2002).
- Tani, H., Morris, R. J. & Kaur, P. Enrichment for murine keratinocyte stem cells based on cell surface phenotype. *Proc. Natl Acad. Sci. USA* **97**, 10960–10965 (2000).
- Li, A., Simmons, P. J. & Kaur, P. Identification and isolation of candidate human keratinocyte stem cells based on cell surface phenotype. *Proc. Natl Acad. Sci. USA* **95**, 3902–3907 (1998).
- Jones, C. *et al.* Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. *Cancer Res.* **64**, 3037–3045 (2004).
- Carter, W. G., Kaur, P., Gil, S. G., Gahr, P. J. & Wayner, E. A. Distinct functions for integrins alpha 3 beta 1 in focal adhesions and alpha 6 beta 4/bullous pemphigoid antigen in a new stable anchoring contact (SAC) of keratinocytes: relation to hemidesmosomes. *J. Cell Biol.* **111**, 3141–3154 (1990).
- Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).
- Li, Y. *et al.* Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells. *Proc. Natl Acad. Sci. USA* **100**, 15853–15858 (2003).
- Goodell, M. A., Brose, K., Paradis, G., Conner, A. S. & Mulligan, R. C. Isolation and functional properties of murine hematopoietic stem cells that are replicating *in vivo*. *J. Exp. Med.* **183**, 1797–1806 (1996).
- Zhou, S. *et al.* The ABC transporter Bcrp1/ABCG2 is expressed in a wide variety of stem cells and is a molecular determinant of the side-population phenotype. *Nature Med.* **7**, 1028–1034 (2001).
- Spangrude, G. J. & Johnson, G. R. Resting and activated subsets of mouse multipotent hematopoietic stem cells. *Proc. Natl Acad. Sci. USA* **87**, 7433–7437 (1990).
- Uchida, N. *et al.* ABC transporter activities of murine hematopoietic stem cells vary according to their developmental and activation status. *Blood* **103**, 4487–4495 (2004).
- Bhattacharya, S. *et al.* Direct identification and enrichment of retinal stem cells/progenitors by Hoechst dye efflux assay. *Invest. Ophthalmol. Vis. Sci.* **44**, 2764–2773 (2003).
- Hierlihy, A. M., Seale, P., Lobe, C. G., Rudnicki, M. A. & Megeney, L. A. The post-natal heart contains a myocardial stem cell population. *FEBS Lett.* **530**, 239–243 (2002).
- Alvi, A. J. *et al.* Functional and molecular characterisation of mammary side population cells. *Breast Cancer Res.* **5**, R1–R8 (2003).
- Smith, G. H. Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* **132**, 681–687 (2005).
- Glimm, H., Oh, I. & Eaves, C. Human hematopoietic stem cells stimulated to proliferate *in vitro* lose engraftment potential during their S/G₂/M transit and do not reenter G₀. *Blood* **96**, 4185–4193 (2000).
- Stingl, J., Emsman, J. T. & Eaves, C. J. in *Methods in Molecular Biology: Basic Cell Culture Protocols* (eds Helgason, C. D. & Miller, C. L.) 249–263 (Humana, New Jersey, 2005).
- Young, L. J. T. in *Methods in Mammary Gland Biology and Breast Cancer Research* (eds Ip, M. M. & Asch, B. B.) 67–74 (Kluwer/Plenum, New York, 2000).
- Szilvassy, S. J., Nicolini, F. E., Eaves, C. J. & Miller, C. L. in *Methods in Molecular Medicine: Hematopoietic Stem Cell Protocols* (eds Jordon, C. T. & Klug, C. A.) 167–187 (Humana, New Jersey, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements This work was supported by grants from the Canadian Stem Cell Network and Genome BC/Canada. J.S. held Postdoctoral Fellowships from the Canadian Breast Cancer Foundation (BC/Yukon Chapter) and the Natural Sciences and Engineering Research Council of Canada; P.E. held a Stem Cell Network Studentship; I.R. and D.C. held British Columbia Cancer Foundation Summer Studentships; and M.S. held a Scholarship from the National Health and Medical Research Council of Australia. The authors thank the Ontario Genomics Innovation Centre for performing the microarrays, the Terry Fox Laboratory FACS Facility for assistance in cell sorting, G. Edin and K. M. Saw for technical assistance, and A. Eaves, S. Aparicio, C. Fisher and D. Kent for scientific discussion.

Author Contributions J.S., P.E., I.R. and D.C. performed most of the experiments described in the manuscript. M.S. and F.V. performed some of the single cell transplants. H.I.L. analysed the microarray data and J.S., P.E. and C.J.E. designed the studies and wrote the manuscript.

Author Information The microarray data can be accessed online at StemBase (<http://www.scbg.ca:8080/StemBase/>) and at the Gene Expression Omnibus (series accession number GSE3711) at <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3711>. Reprints and permissions information is available at www.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and request for materials should be directed to C.J.E. (ceaves@bccrc.ca).

LETTERS

Eisosomes mark static sites of endocytosis

Tobias C. Walther^{1*}, Jason H. Brickner^{1*†}, Pablo S. Aguilar¹, Sebastián Bernales¹, Carlos Pantoja¹ & Peter Walter¹

Endocytosis functions to recycle plasma membrane components, to regulate cell-surface expression of signalling receptors and to internalize nutrients in all eukaryotic cells. Internalization of proteins, lipids and other cargo can occur by one of several pathways that have different, but often overlapping, molecular requirements^{1–5}. To mediate endocytosis, effectors assemble transiently underneath the plasma membrane, carry out the mechanics of membrane deformation, cargo selection and vesicle internalization, and then disassemble. The mechanism by which endocytosis initiates at particular locations on the plasma membrane has remained unknown. Sites of endocytosis might be formed randomly, induced by stochastic protein and/or lipid clustering. Alternatively, endocytosis might initiate at specific locations. Here we describe large immobile protein assemblies at the plasma membrane in the yeast *Saccharomyces cerevisiae* that mark endocytic sites. These structures, termed eisosomes (from the Greek 'eis', meaning into or portal, and 'soma', meaning body), are composed primarily of two cytoplasmic proteins, Pil1 and Lsp1. A plasma membrane protein, Sur7, localizes to eisosomes. These structures colocalize with sites of protein and lipid endocytosis, and their components genetically interact with known endocytic effectors. Loss of Pil1 leads to clustering of eisosome remnants and redirects endocytosis and endocytic effector proteins to these clusters.

Examination of a library of yeast proteins tagged with green fluorescent protein (GFP) identified two closely related cytoplasmic proteins, Pil1 and Lsp1, that share 72% sequence identity and localize in a marked pattern at the cell cortex⁶ (Fig. 1). The amino acid sequences of Lsp1 and Pil1 do not contain any recognizable functional domains. Fluorescence and confocal microscopy of cells expressing GFP-tagged Pil1 or Lsp1 showed 20–45 bright dots, termed 'eisosomes', at the cell periphery underneath the plasma membrane (Fig. 1a and Supplementary Information). Fluorescently tagged Pil1 and Lsp1 colocalize in these structures (Fig. 1c). Moreover, co-immunoprecipitation showed that Pil1 and Lsp1 are the main proteins in these complexes and are present in roughly equimolar proportions (Fig. 1b).

To find additional eisosome components, we further inspected the yeast GFP database and identified a previously characterized family of structurally related transmembrane proteins with a similar localization pattern^{6–8}: Sur7 (Fig. 1a), Ydl222c and Ynl194c. Further analysis showed colocalization of Sur7–GFP with Pil1 tagged with the haemagglutinin A (HA) epitope (Fig. 1d).

Quantification of the fluorescence intensity of individual eisosomes revealed a tight distribution of GFP-tagged molecules per eisosome (Fig. 1a, bottom). Thus, different eisosomes contain similar numbers (within a factor of two) of constituent molecules. As estimated by quantitative western blot of tandem affinity purification (TAP)-tagged proteins, there are roughly 100,000 molecules of both

Pil1 and Lsp1 per cell⁹. We detected no perceptible pool of soluble Pil1–GFP. We therefore estimate that each eisosome comprises about 2,000–5,000 molecules of each protein.

To test whether eisosomes show dynamic behaviour, we observed single cells in time-lapse movies. Eisosomes are unexpectedly stable structures that do not disassemble or move noticeably even over the 30-min time course of the experiment (Fig. 1e). Eisosomes also do not exchange subunits with a cytoplasmic pool: fluorescence recovery after photobleaching (FRAP) analysis of Pil1–GFP-labelled eisosomes showed no detectable recovery of fluorescence to bleached eisosomes over a 20-min time course (Fig. 1f, g). Thus, eisosomes are stable ultrastructural assemblies in cells that, akin to centrosomes, could be considered as unique organelles.

SUR7 has been previously characterized as an overexpression suppressor of the conditional growth defect of a null mutation in *RVS161*, encoding the yeast homologue of mammalian endophilin, an endocytic effector^{8,10}. This observation suggested to us that eisosomes might have a role in endocytosis. We therefore monitored endocytosis of HA-tagged Ste3 (the α -factor mating pheromone receptor) by following its degradation after delivery to the vacuole¹¹. In this assay, deletion of *PIL1* ($t_{1/2} = 27 \pm 3$ min), *LSP1* ($t_{1/2} = 24 \pm 3$ min) or both ($t_{1/2} = 32 \pm 5$ min) substantially decreased the rate of endocytosis compared with wild-type cells ($t_{1/2} = 18 \pm 2$ min; see Supplementary Information).

To test whether *PIL1* and *LSP1* interact genetically with *RVS161*, we monitored endocytosis of Ste3 in strains lacking these genes in combination. In agreement with previous work¹¹, we observed a strong endocytosis defect in cells deleted for *RVS161* ($t_{1/2} = 112 \pm 37$ min). This defect was almost completely suppressed by deletion of *LSP1* ($t_{1/2} = 30 \pm 8$ min), suggesting that Lsp1 negatively regulates endocytic function in the *rvs161* Δ mutant. This suppression requires Pil1 because strains lacking all three components show no suppression ($t_{1/2} = 74 \pm 10$ min; not significantly different from *rvs161* Δ). Together, the data show that *LSP1* and *RVS161* interact genetically and support the notion that Pil1 and Lsp1 affect endocytosis, having overlapping but separable functions.

Deletion of *RVS161* can also be suppressed by deletion of genes involved in sphingolipid biosynthesis, suggesting that endocytosis can be promoted by perturbing sphingolipid abundance in the cell¹². Moreover, Pil1 and Lsp1 have been described as substrates of two sphingolipid-regulated kinases, Pkh1 and Pkh2 (ref. 13), and mutations in these kinases^{14,15} or in sphingolipid biosynthetic enzymes¹⁶ block endocytosis. To determine whether loss of Pil1 or Lsp1 affects endocytosis by altering the lipid composition of the cell, we determined the relative abundance of ceramides and long-chain bases in *pil1* Δ , *lsp1* Δ and *pil1* Δ *lsp1* Δ strains and found that they were not measurably altered (see Supplementary Information). Therefore, the genetic interactions between *PIL1*, *LSP1* and *RVS161* described

¹Howard Hughes Medical Institute and Department of Biochemistry and Biophysics, University of California at San Francisco, 600-16th Street, San Francisco, California 94143-2200, USA. [†]Present address: Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60208, USA.

*These authors contributed equally to this work.

above do not indirectly result from effects on the lipid pools in the mutant cells.

Microscopic analyses of *pil1Δ* cells showed that Pil1 is required for proper localization of both Lsp1–GFP (Fig. 2a) and Sur7–GFP (Fig. 2b): in the absence of Pil1, Lsp1–GFP localized to only a few bright clusters (less than four per cell) at the cell periphery and to a diffusely staining cytoplasmic pool. Analogously, in the absence of Pil1, Sur7 localized to a few bright clusters and a pool diffusely distributed over the plasma membrane. By contrast, deletion of *LSP1* did not affect the localization of Sur7–GFP (Fig. 2b) or Pil1–GFP (not shown), and deletion of *SUR7* (alone or in combination with its orthologues Ydl222c and Ynl194c) did not affect the localization of Lsp1–GFP or Pil1–GFP (data not shown). These results identify Pil1 as the fundamental component of eisosomes required for their construction or maintenance and further support the notion that, despite their high degree of sequence identity, Pil1 and Lsp1 have distinct nonredundant functions.

To study the relationship of eisosomes to sites of endocytosis, we monitored the uptake of the fluorescent dye FM4-64, a lipid marker

for endocytosis^{1,3,4}, in cells expressing Pil1–GFP. We labelled cells with FM4-64 and blocked endocytosis either during labelling or shortly after labelling by poisoning the cells with azide and fluoride on ice as described¹. As expected, when endocytosis was blocked during labelling (Fig. 3a, $t = 0$), we observed FM4-64 staining of only the plasma membrane. When endocytosis was blocked after a 20-s incubation at room temperature, the dye concentrated into discrete foci, representing early endocytic intermediates¹⁷. Strikingly, all of the FM4-64-labelled foci localized to eisosomes (Fig. 3a), although not all eisosomes nucleated FM4-64 foci.

Movies following FM4-64 uptake in living cells also showed the formation of FM4-64 foci that always colocalized with eisosomes. As expected for *bona fide* endocytosis intermediates, FM4-64 labelling of eisosomes was transient and consumed within 5–10 min as the dye was transported to endosomal and vacuolar membranes (Fig. 3b, 600 s). Thus, at any given time, only a subset of eisosomes are active for endocytosis as monitored by this method. Monitoring the time course of FM4-64 internalization in strains bearing the *pil1Δ*, *lsp1Δ* or *rvs161Δ* mutation either alone or in combination showed kinetic delays of FM4-64 uptake consistent with the delays observed for Ste3 internalization (data not shown).

To test whether eisosomes also colocalize with sites of protein uptake, we made use of the regulated endocytosis of the hexose transporter Hxt2. Cells grown in low-glucose media express Hxt2 and transport it to the plasma membrane. On shift to high-glucose media, Hxt2 synthesis is arrested and the protein is removed from the plasma membrane by endocytosis¹⁸. Similar to FM4-64, Hxt2–GFP accumulated in discrete foci on the plasma membrane. The foci were consumed as Hxt2 was delivered to endosomes and then to vacuoles. However, endocytosis of Hxt2–GFP occurred more slowly than that of FM4-64 (Fig. 3d). Thus, to determine whether these presumed endocytic intermediates were the same, we added FM4-64 at different times after shift to high glucose. Under such conditions, all Hxt2–GFP foci colocalized with early FM4-64 endocytic intermediates (Fig. 3d, left), and thus with eisosomes. Together, these results indicate that eisosomes have a role in endocytosis of both protein and lipid cargos.

If eisosomes mark sites of endocytosis, then their redistribution would be expected to change the sites of endocytosis. To test this prediction, we made use of the clustering of eisosomal remnants

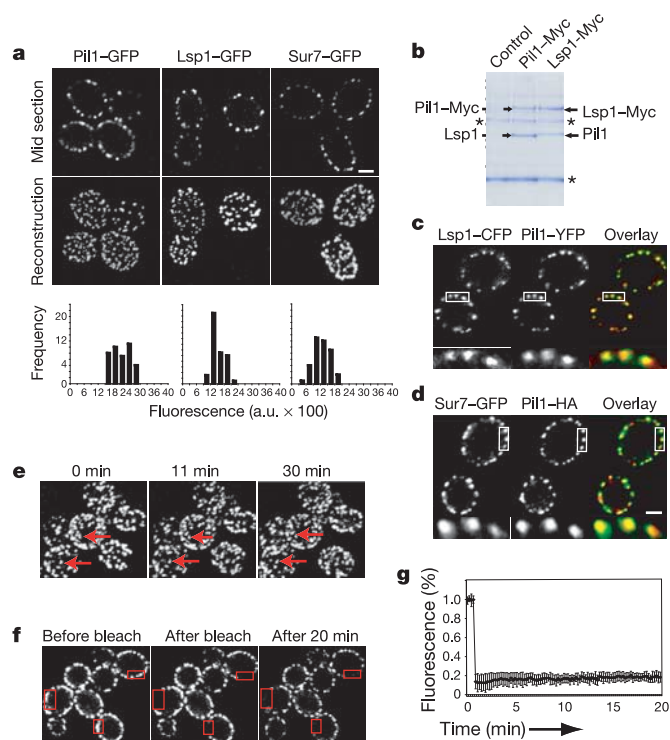


Figure 1 | Eisosomes are large protein assemblies at the cell cortex. **a**, Pil1, Lsp1 and Sur7 colocalize to patches beneath the plasma membrane. GFP-tagged Pil1, Lsp1 and Sur7 were localized in live cells by acquiring confocal z-stacks. Top, mid section and projection of a three-dimensional reconstruction of the cell. Bottom, distribution of the fluorescence intensities of Pil1–GFP, Lsp1–GFP and Sur7–GFP patches plotted in classes (y axis, number of spots) of arbitrary fluorescence units (x axis, 12 bit). **b**, Pil1 and Lsp1 interact physically. Myc-tagged Lsp1 and Pil1 were immunoprecipitated from cell extracts and labelled bands were identified by mass spectrometry. Asterisks indicate IgG heavy and light chains. **c**, Pil1 and Lsp1 colocalize in live cells. A representative mid section is shown. Insets show magnification of the boxed region. **d**, Sur7 colocalizes with Pil1. Sur7–GFP and Pil1–HA were visualized by indirect immunofluorescence and confocal microscopy. A representative mid section is shown. Scale bar, 2 μ m. **e**, Eisosomes are immobile. Cells expressing Pil1–GFP cells were followed by time-lapse confocal microscopy for 30 min. Projections of three-dimensional reconstructions of cells at 0, 11 and 30 min are shown. Arrows mark two eisosomes as visual reference points. **f**, FRAP of Pil1–GFP. Images of cells before, immediately after bleaching, and after 20 min of recovery are shown. Bleached areas are boxed. **g**, Quantification of fluorescence recovery over time averaged between three independent experiments.

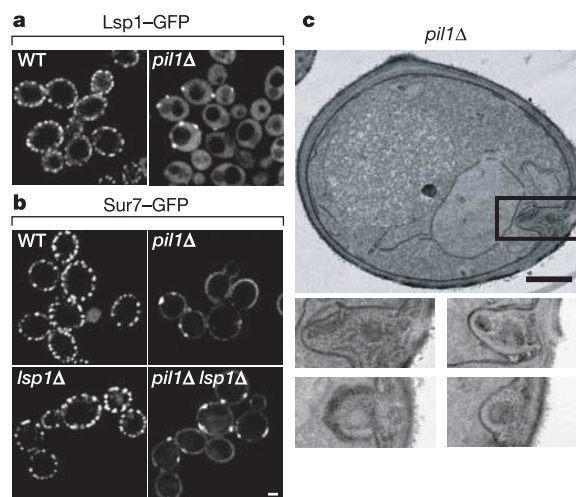


Figure 2 | Pil1 is required for Lsp1 and Sur7 localization. **a**, Confocal image of Lsp1–GFP expressed in wild-type and *pil1Δ* cells. A representative field of mid sections is shown. **b**, Confocal images of Sur7–GFP expressed in wild-type, *pil1Δ*, *lsp1Δ* and *pil1Δ lsp1Δ* strains. Scale bar, 2 μ m. **c**, Top panel, electron micrograph of *pil1Δ* cells fixed with glutaraldehyde and stained with uranyl acetate and permanganate. Bottom panels, magnified large aberrant membrane invaginations seen only in mutant cells. Scale bar, 0.5 μ m.

observed in *pil1Δ* strains (Fig. 2). Notably, after FM4-64 uptake in *pil1Δ* cells expressing Sur7-GFP, we observed colocalization of all early FM4-64 foci with the clusters (Fig. 3c). Similarly, foci defining Hxt2-GFP endocytosis were redirected to the clusters in *pil1Δ* cells (Fig. 3d, right).

Analysis of *pil1Δ* cells by electron microscopy revealed large, aberrant plasma membrane invaginations (Fig. 2c) not observed in wild-type cells. From their frequency of appearance in the cell sections, we estimate that every *pil1Δ* cell contains at least one such structure (Supplementary Information), consistent with the frequency of clustering of endocytic sites found in *pil1Δ* cells by FM4-64 and Hxt2-GFP fluorescence. Thus, the aberrant plasma membrane invaginations seen in the electron micrographs are likely to be sites of clustered eisosome remnants. These results suggest that eisosomes are required for the proper spatial distribution of endocytic events. In the absence of Pil1, the remaining eisosomal components collapse into clusters, representing large and aberrantly structured endocytic sites. Endocytosis at the aberrant sites functions with reduced capacity but is efficient enough to sustain cell viability.

The mechanics of endocytosis are carried out by effector proteins that assemble at the membrane, where they mediate vesicle

formation, pinching-off, and movement into the cell's interior. Several parallel endocytic pathways are thought to use subsets of endocytic effectors, may select different cargos, and, indeed, may be localized to discrete domains of the cell surface². Genetic and microscopic analyses indicate that the transient formation of actin patches at the cell cortex is an important step in the process². Mutations in numerous genes encoding cortical actin patch components lead to defects in endocytosis², and most of these proteins have homologues in mammalian cells, where they likewise participate in endocytosis. The dynamic assembly of actin patches at the cell cortex is thought to promote the formation of endocytic vesicles by propelling them from the plasma membrane into the cell interior through a defined coordinated sequence of molecular events.

The sequential recruitment of endocytic effectors has been visualized directly^{19,20}. The assembly occurs in several phases, each of which is marked by the transient (~20–30 s) presence of specific actin patch proteins. Colocalization of one such protein, Abp1, with FM4-64 in early endocytic intermediates showed that these patches are indeed sites of endocytosis¹⁷. To examine the relationship between eisosomes and cortical actin patches, we first tested the effects of latrunculin A (Lat A), a drug that prevents actin

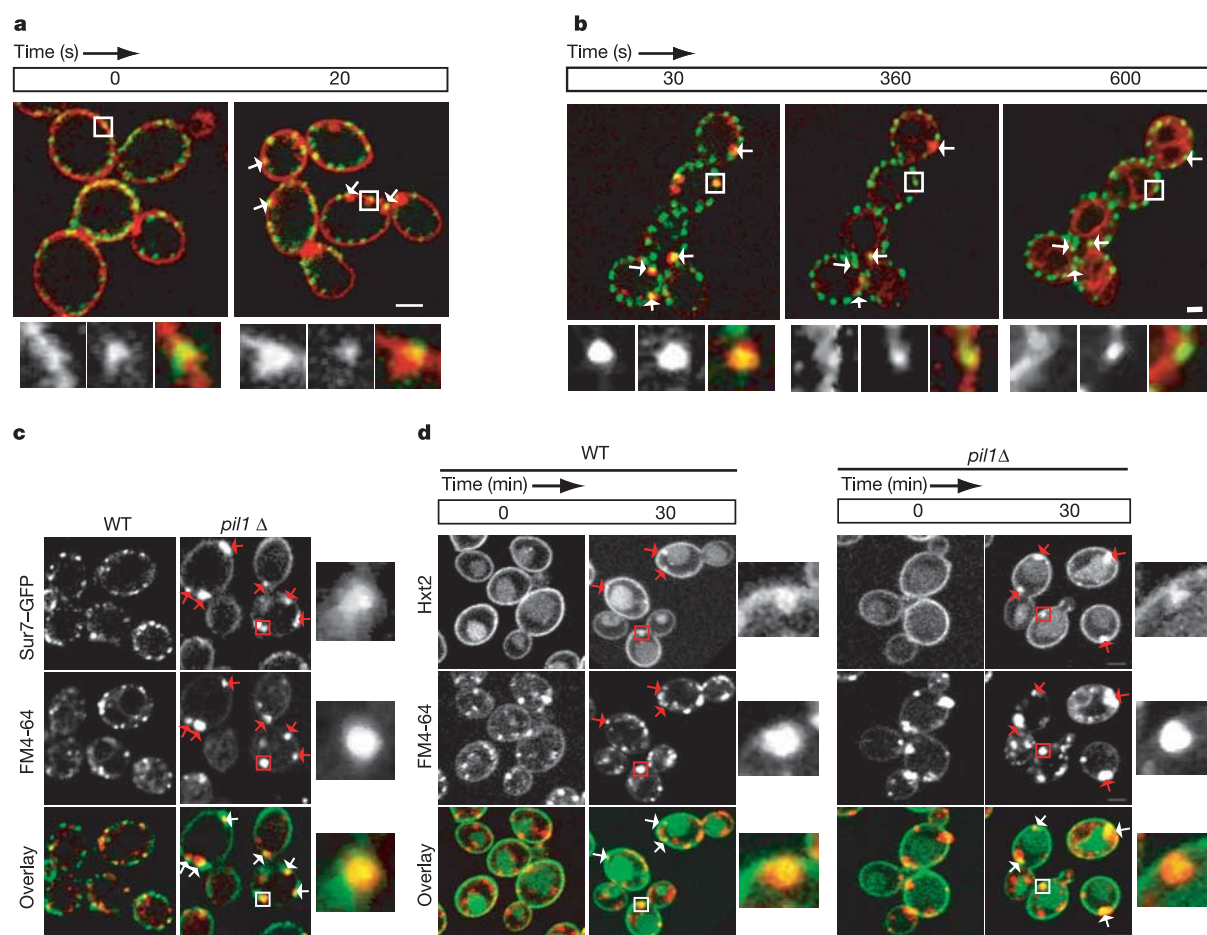


Figure 3 | Eisosomes colocalize with sites of endocytosis. **a**, Cells expressing Pil1-GFP were labelled with FM4-64 on ice. To remove excess dye, cells were washed with ice-cold YPD and subsequently resuspended in room-temperature YPD. Endocytosis was stopped by transfer to ice and the addition of 10 mM Na₃N and 10 mM NaF either immediately or after 20 s. Top, confocal microscopy images. Bottom, magnification of single eisosomes (boxed above), showing, from left to right, FM4-64 staining, GFP staining and overlay. Arrows indicate representative examples of FM4-64 and Pil1-GFP colocalization. Scale bar, 2 μ m. **b**, Cells labelled as in **a** and imaged at the indicated times after warming to room temperature. **c**, Endocytosis can be redirected to mislocalized eisosome remnants. Uptake

of FM4-64 was visualized in wild-type or *pil1Δ* cells expressing Sur7-GFP. Arrows indicate representative examples of FM4-64 and Sur7-GFP colocalization. Right images show close-ups of a representative eisosome (boxed). **d**, Hxt2 is taken up through eisosomes. Wild-type or *pil1Δ* cells expressing Hxt2-GFP were grown in 0.1% glucose for 6 h and then shifted to 5% glucose. Uptake of Hxt2 was much slower than that of FM4-64; thus, the number of foci at any given time was lower. At the indicated time points, cells were pulse-labelled with FM4-64 and imaged after a 30-s pulse to catch early FM4-64 endocytic intermediates. Arrows indicate representative examples of FM4-64 and Hxt2-GFP foci colocalization. Right images show close-ups of a representative (boxed) eisosome (WT) or eisosome remnant (*pil1Δ*).

polymerization and inhibits endocytosis²¹. Lat A did not block the concentration of FM4-64 at eisosomes, but did prevent Abp1-GFP from assembling into cortical actin patches (Fig. 4a), thus indicating that the Lat A treatment was effective. In contrast to untreated cells, however, FM4-64 foci in Lat-A-treated cells were not consumed even at late time points. Together, these results show that the formation of FM4-64 foci at the plasma membrane is actin-independent, but that their consumption, namely the depletion of FM4-64 from eisosomes, is actin-dependent¹⁷.

Imaging of both Abp1-GFP and FM4-64 revealed two populations of actin patches: one that colocalized with FM4-64 foci (and thus with eisosomes) and another that appeared at unrelated positions near the cell cortex. To confirm that Abp1 patches colocalizing with FM4-64 were functionally related to eisosomes, we observed Abp1 patches in *pil1Δ* mutant cells in which eisosome remnants are clustered. As expected, Abp1 patches formed in *pil1Δ* mutant cells more frequently at the clustered sites (Fig. 4b and Supplementary Information). Owing to crowding at the cell cortex, images of both eisosome and non-eisosome-localized Abp1 patches were difficult to interpret. We therefore exploited the transient nature of the actin patches and integrated the Abp1-GFP signal over time. A clear peak of the Abp1-GFP-containing patches coinciding with clustered eisosome remnants emerged over time, whereas the population of patches that appeared at unrelated and apparently random positions became averaged into a continuous rim at the cell periphery (Fig. 4b). Another actin patch protein, Sla2, showed a very similar behaviour (Fig. 4c).

Our results show that actin patches form at eisosomes, but Abp1- and Sla2-containing actin patches also formed at unrelated sites at the cortex. Furthermore, whereas Pil1-GFP labelled eisosomes are not visible in small buds (T.C.W. and P.W., unpublished data), Abp1- and Sla2-containing patches formed frequently in small buds (see Fig. 4c). Assuming that formation of these actin patches in each case marks productive endocytic events, they probably represent eisosome-independent endocytic events that are not monitored by FM4-64 uptake. Actin patches in buds, for example, did not collapse into clusters in *pil1Δ* cells (Fig. 4c), indicating that eisosome-mediated endocytosis may be especially important in mother cells. Thus, like metazoan cells, yeast may have parallel endocytic pathways that have different molecular requirements. According to this view, one

pathway, localized to eisosomes and defined by its sensitivity to eisosome disruption, primarily operates in the mother cell, whereas others, less sensitive to eisosome disruption, primarily operate in the bud. These pathways share at least some common components, such as the actin patch proteins analysed here.

The notion that endocytic events initiating at eisosomes use actin patch proteins is further underscored by genetic interactions between eisosome components and endocytic effectors. For example, *pil1Δ* mutants show synthetic lethality with a conditional allele of *PAN1* (*pan1-20*) and synthetic growth defects with *sla1Δ* mutations (Supplementary Information). Such synthetic defects could indicate that two components closely interact. Alternatively, each mutation could compromise alternative but redundant pathways: when both mutations are combined, reciprocal compensation between the pathways would be lost. We also observed suppression of *end4-1* (*END4* is allelic with *SLA2*) mutants by deletion of *PIL1* and suppression of the endocytic defect of *rvs161Δ* by deletion of *LSP1*. Thus, although so far the complete matrix of possible interactions has been only explored superficially, there is an extensive web of genetic, biochemical and cytological interactions that solidly ties the three known eisosome components to the endocytic machinery (Fig. 4d).

Taken together, the notion that eisosomes have a role in endocytosis is supported by the following observations: first, early endocytic intermediates form initially at eisosomes; second, these are built in an actin-independent, but consumed in an actin-dependent, reaction; third, a subpopulation of cortical actin patches colocalizes with eisosomes; fourth, endocytosis of both lipid and protein cargo, and cortical actin patches can be redirected to clustered eisosome remnants in *pil1Δ* strains; fifth, *pil1Δ* *lsp1Δ* mutant strains are impaired in endocytosis and show large, aberrant plasma membrane invaginations; and sixth, *LSP1* and *PIL1* connect through a web of genetic and biochemical interactions with known endocytic effectors, such as *RVS161*, *RVS167*, *PAN1*, *SLA1* and *SLA2*. Eisosomes consist of two main cytoplasmic components, Lsp1 and Pil1, the plasma membrane protein Sur7, and perhaps other undefined proteins, such as other members of the Sur7 protein family. By contrast to other previously described endocytosis components, eisosomes are immobile static structures. Eisosomes thus provide a stable, functional link between the cytoplasmic components that

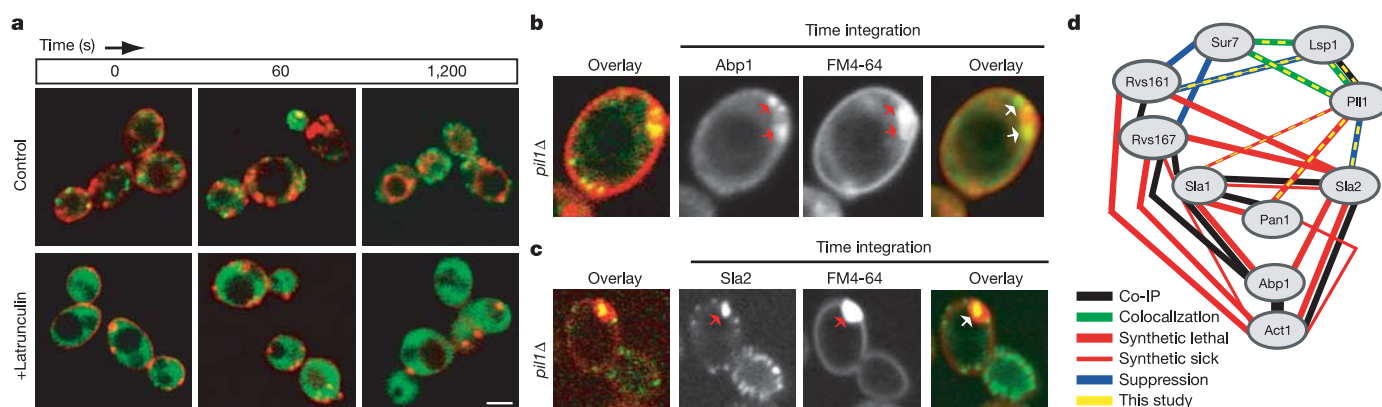


Figure 4 | Endocytic actin patches form at eisosomes. **a**, Formation of early endocytic intermediates is actin-independent. Cells expressing Abp1-GFP were treated with or without Lat A, pulse-labelled with FM4-64, incubated at room temperature for the indicated times, and imaged by confocal microscopy. **b**, Confocal time-lapse images of *pil1Δ* cells expressing Abp1-GFP and pulsed with FM4-64. An individual still frame of this time series is shown on the left. Three images integrated over a 4-min observation period are shown on the right. Note that the small sharp foci of Abp1-GFP staining in the still frame have disappeared in the time-averaged images because they did not recur at the same loci. By contrast, the Abp1-GFP signal at the eisosome cluster amplified during time-averaging (arrows).

c, *pil1Δ* cells expressing Sla2-GFP treated as in **b**. Note the frequent appearance of Sla2-containing actin patches in the bud (lower right). These patches do not relocalize to clustered eisosome remnants as observed in the mother cells and may represent a parallel, eisosome-independent pathway. **d**, Network of interactions of eisosome components with known endocytic effectors. Physical and genetic and biochemical interactions and colocalization are indicated by colour coding. Data for interactions reported in this study (dashed yellow lines superimposed) and references for previously established interactions are given in the Supplementary Information.

catalyse endocytosis and the plasma membrane, and may regulate when and where endocytosis occurs.

Many observations point to an eisosome–lipid connection. First, Sur7 has been suggested to function in endocytosis because it is an overexpression suppressor of a deletion of *RVS161* and *RVS167*, encoding yeast endophilin and amphiphysin, respectively¹⁰. In metazoan cells, these proteins are thought to modulate membrane curvature after invagination of endocytic pits. Other suppressors of *rvs161Δ* map to genes involved in sphingolipid biosynthesis. Second, the lipid-regulated kinases Pkh1 and Pkh2 phosphorylate Pil1 and Lsp1 (ref. 13). Pkh2 physically associates with Pil1 and Lsp1 and shows an eisosome-like localization^{22,23}. Last, long-chain bases, the precursors of sphingolipids, inhibit the phosphorylation of Pil1 but stimulate the phosphorylation of Lsp1. Thus, differential regulation of Pil1 and Lsp1 could modulate the activity of eisosomes to recruit other endocytosis components and concentrate them beneath the plasma membrane where they participate in the mechanics of membrane uptake by endocytosis, perhaps ascertaining that the local lipid composition is compatible with a productive endocytic event. In the *PIL1* and *LSP1* mutants, a diminished local concentration of endocytosis effectors would result in a reduced rate of uptake. Their differential regulation by phosphorylation parallels the opposite effects that *PIL1* and *LSP1* deletions have on endocytosis in the context of *rvs161Δ*. Thus, Lsp1 could be a negative regulator or ‘governor’, restricting endocytosis to a few sites. Its deletion would then increase endocytosis under suboptimal conditions, providing a plausible explanation for its genetic interaction with the *rvs161Δ* mutation.

In many metazoan cell types, endocytosis and exocytosis are spatially segregated into mutually exclusive zones and both can occur repeatedly at the same sites^{24,25}. In synaptic nerve terminals, endocytic activity is organized in membrane patches forming a ring surrounding the central zone of exocytosis, enabling the cell to coordinate efficiently bursts of exocytic activity with endocytic membrane recycling²⁶. Endocytic effectors including endophilin, synaptojanin and AP180 are concentrated in the endocytic zone by scaffold proteins, such as Dap160 (also known as intersectin). In *Drosophila* mutants lacking intersectin, these effector proteins are no longer concentrated around the active zone. Intriguingly, like *pil1Δ* mutants, these mutant cells lacking intersectin also show only a modest kinetic delay in endocytosis, even though the structural organization of the endocytic zones is severely damaged²⁷. Putative homologues of Pil1 and Lsp1 (ref. 13) and potentially other eisosome components might function similarly to organize endocytosis in higher eukaryotes.

A mechanism to regulate endocytosis spatially—that is, to choose which of several possible sites of endocytosis to activate—is likely to be important for maintaining organization of the plasma membrane. In the absence of such regulation, plasma membrane structure can be grossly affected, as seen here by the large aberrant invaginations formed in *pil1Δ* cells. A controlled imbalance between the two processes through local enhancement or inhibition of endocytosis could also be used to localize cell growth and to contribute to the establishment and maintenance of cell polarity. Eisosomes may thus have a profound role in regulating plasma membrane architecture.

METHODS

Yeast strains and plasmids. All yeast strains were derived from the W303 strain by using PCR-based modification²⁸.

Protein analyses. To immunopurify Pil1 or Lsp1 complexes, 50 ml of cells expressing Myc-tagged proteins were lysed in the presence of 1% Triton X-100 (150 mM NaCl and 20 mM HEPES; pH 7.4), incubated with agarose conjugated to anti-Myc antibodies for 2 h, washed, and eluted in 2% SDS. Proteins were identified by mass fingerprinting after SDS–PAGE using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.

Endocytosis assays. We monitored endocytosis by uptake of FM4-64 (Molecular Probes) as described¹. In short, cells were labelled on ice (40 μM), washed three times with ice-cold YPD medium and resuspended at *t* = 0 in YPD (room

temperature). Samples were either stopped by the addition of 10 mM Na₃N and 10 mM NaF or imaged continuously. For Lat A treatment, 50 μM was added for 15 min before start of the assay; Lat A was also included in subsequent media.

Microscopy. Light microscopy was done on an LSM510 confocal microscope (Zeiss). Images were processed using LSM (Zeiss) and ImageJ (W. S. Rasband; <http://rsb.info.gov/ij/>) software. Immunofluorescence and electron microscopy were done as described^{29,30}.

Received 12 April; accepted 16 November 2005.

- Vida, T. A. & Emr, S. D. A new vital stain for visualizing vacuolar membrane dynamics and endocytosis in yeast. *J. Cell Biol.* **128**, 779–792 (1995).
- Engqvist-Goldstein, A. E. & Drubin, D. G. Actin assembly and endocytosis: from yeast to mammals. *Annu. Rev. Cell. Dev. Biol.* **19**, 287–332 (2003).
- Zheng, B., Wu, J. N., Schober, W., Lewis, D. E. & Vida, T. Isolation of yeast mutants defective for localization of vacuolar vital dyes. *Proc. Natl Acad. Sci. USA* **95**, 11721–11726 (1998).
- Wendland, B., McCaffery, J. M., Xiao, Q. & Emr, S. D. A novel fluorescence-activated cell sorter-based screen for yeast endocytosis mutants identifies a yeast homologue of mammalian eps15. *J. Cell Biol.* **135**, 1485–1500 (1996).
- Perret, E., Lakkaraju, A., Deborde, S., Schreiner, R. & Rodriguez-Boulán, E. Evolving endosomes: how many varieties and why? *Curr. Opin. Cell Biol.* **17**, 423–434 (2005).
- Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
- Malinska, K., Malinsky, J., Opekarova, M. & Tanner, W. Distribution of Can1p into stable domains reflects lateral protein segregation within the plasma membrane of living *S. cerevisiae* cells. *J. Cell Sci.* **117**, 6031–6041 (2004).
- Young, M. E. *et al.* The Sur7p family defines novel cortical domains in *Saccharomyces cerevisiae*, affects sphingolipid metabolism, and is involved in sporulation. *Mol. Cell. Biol.* **22**, 927–934 (2002).
- Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
- Sivados, P., Peypouquet, M. F., Doignon, F., Aigle, M. & Crouzet, M. Cloning of the multicopy suppressor gene *SUR7*: evidence for a functional relationship between the yeast actin-binding protein Rvs167 and a putative membranous protein. *Yeast* **13**, 747–761 (1997).
- Munn, A. L., Stevenson, B. J., Geli, M. I. & Riezman, H. *end5*, *end6*, and *end7*: mutations that cause actin delocalization and block the internalization step of endocytosis in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **6**, 1721–1742 (1995).
- Dickson, R. C. & Lester, R. L. Metabolism and selected functions of sphingolipids in the yeast *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta* **1438**, 305–321 (1999).
- Zhang, X., Lester, R. L. & Dickson, R. C. Pil1p and Lsp1p negatively regulate the 3-phosphoinositide-dependent protein kinase-like kinase Pkh1p and downstream signaling pathways Pkc1p and Ypk1p. *J. Biol. Chem.* **279**, 22030–22038 (2004).
- Friant, S., Lombardi, R., Schmelzle, T., Hall, M. N. & Riezman, H. Sphingoid base signaling via Pkh kinases is required for endocytosis in yeast. *EMBO J.* **20**, 6783–6792 (2001).
- deHart, A. K., Schnell, J. D., Allen, D. A. & Hicke, L. The conserved Pkh–Ypk kinase cascade is required for endocytosis in yeast. *J. Cell Biol.* **156**, 241–248 (2002).
- Zanolari, B. *et al.* Sphingoid base synthesis requirement for endocytosis in *Saccharomyces cerevisiae*. *EMBO J.* **19**, 2824–2833 (2000).
- Huckaba, T. M., Gay, A. C., Pantaleón, L. F., Yang, H. C. & Pon, L. A. Live cell imaging of the assembly, disassembly, and actin cable-dependent movement of endosomes and actin patches in the budding yeast, *Saccharomyces cerevisiae*. *J. Cell Biol.* **167**, 519–530 (2004).
- Kruckeberg, A. L., Ye, L., Berden, J. A. & van Dam, K. Functional expression, quantification and cellular localization of the Hxt2 hexose transporter of *Saccharomyces cerevisiae* tagged with the green fluorescent protein. *Biochem. J.* **339**, 299–307 (1999).
- Kaksonen, M., Sun, Y. & Drubin, D. G. A pathway for association of receptors, adaptors, and actin during endocytic internalization. *Cell* **115**, 475–487 (2003).
- Kaksonen, M., Toret, C. P. & Drubin, D. G. A modular design for the clathrin- and actin-mediated endocytosis machinery. *Cell* **123**, 305–320 (2005).
- Ayscough, K. R. *et al.* High rates of actin filament turnover in budding yeast and roles for actin in establishment and maintenance of cell polarity revealed using the actin inhibitor latrunculin-A. *J. Cell Biol.* **137**, 399–416 (1997).
- Roelants, F. M., Torrance, P. D., Bezman, N. & Thorner, J. Pkh1 and pkh2 differentially phosphorylate and activate ypk1 and ykr2 and define protein kinase modules required for maintenance of cell wall integrity. *Mol. Biol. Cell* **13**, 3005–3028 (2002).
- Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
- Gaidarov, I., Santini, F., Warren, R. A. & Keen, J. H. Spatial control of coated-pit dynamics in living cells. *Nature Cell Biol.* **1**, 1–7 (1999).
- Gundelfinger, E. D., Kessels, M. M. & Qualmann, B. Temporal and spatial coordination of exocytosis and endocytosis. *Nature Rev. Mol. Cell. Biol.* **4**, 127–139 (2003).

26. Roos, J. & Kelly, R. B. The endocytic machinery in nerve terminals surrounds sites of exocytosis. *Curr. Biol.* **9**, 1411–1414 (1999).
27. Marie, B. *et al.* Dap160/intersectin scaffolds the periactive zone to achieve high-fidelity endocytosis and normal synaptic growth. *Neuron* **43**, 207–219 (2004).
28. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14**, 953–961 (1998).
29. Brickner, J. H. & Walter, P. Gene recruitment of the activated *INO1* locus to the nuclear membrane. *PLoS Biol* **2**, e342 (2004).
30. Heiman, M. G. & Walter, P. Prm1p, a pheromone-regulated multispinning membrane protein, facilitates plasma membrane fusion during yeast mating. *J. Cell Biol.* **151**, 719–730 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Kelly, G. Davis, M. von Zastrow, S. Emr, D. Drubin and members of the Drubin and Walter laboratories for discussion and comments; N. Davis for the Ste3-HA plasmid; and D. Brickner for technical assistance. T.C.W. was supported by the Human Frontier Science Program, J.H.B. by the Helen Hay Whitney Foundation, P.S.A. by the Damon Runyon Cancer Research Fund and Fundación Antorchas, and S.B. by the UCSF Herbert H. Boyer Fund. This work was supported in part by a grant from the NIH and by a Sandler Opportunity Fund Award (UCSF). P.W. is an Investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.C.W. (twalth@itsa.ucsf.edu).

LETTERS

A *var* gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria

Till S. Voss¹, Julie Healer¹, Allison J. Marty^{1,2}, Michael F. Duffy³, Jennifer K. Thompson¹, James G. Beeson¹, John C. Reeder⁴, Brendan S. Crabb¹ & Alan F. Cowman¹

Mono-allelic expression of gene families is used by many organisms to mediate phenotypic variation of surface proteins. In the apicomplexan parasite *Plasmodium falciparum*, responsible for the severe form of malaria in humans, this is exemplified by antigenic variation of the highly polymorphic *P. falciparum* erythrocyte membrane protein 1 (PfEMP1)^{1,2}. PfEMP1, encoded by the 60-member *var* gene family^{3–6}, represents a major virulence factor due to its central role in immune evasion and intravascular parasite sequestration. Mutually exclusive expression of PfEMP1 is controlled by epigenetic mechanisms involving chromatin modification and perinuclear *var* locus repositioning^{7,8}. Here we show that a *var* promoter mediates the nucleation and spreading of stably inherited silenced chromatin. Transcriptional activation of this promoter occurs at the nuclear periphery in association with chromosome-end clusters. Additionally, the *var* promoter sequence is sufficient to infiltrate a transgene into the allelic exclusion programme of *var* gene expression, as transcriptional activation of this transgene results in silencing of endogenous *var* gene transcription. These results show that a *var* promoter is sufficient for epigenetic silencing and mono-allelic transcription of this virulence gene family, and are fundamental for our understanding of antigenic variation in *P. falciparum*. Furthermore, the PfEMP1 knockdown parasites obtained in this study will be important tools to increase our understanding of *P. falciparum*-mediated virulence and immune evasion.

In individual *P. falciparum* parasites one *var* gene is expressed and switching to an alternative occurs by *in situ* transcriptional activation^{9,10}. *var* genes are flanked by one of three upstream sequences (*upsA*, *upsB* and *upsC*)⁶, and differential expression of these subtypes is linked to disease¹¹. *UpsA*- and *upsB*-type *var* genes are located subtelomerically, whereas *upsC*-type *var* genes are in chromosome-internal clusters. *UpsB* and *upsC* sequences display promoter activities^{12,13} and interact with DNA-binding proteins¹⁴. Cooperative interactions *in cis* between *upsC* and the *var* intron have been implicated in silencing^{15–17}. A role for silent information regulator 2 (PfsIR2) in *var* silencing has been demonstrated^{7,8}. These findings suggest a vital function of *var* promoters in epigenetic regulation of this family.

To test this hypothesis we transfected *P. falciparum* strain 3D7 with plasmids pHBupsC, pHBupsC^R and pHBupsC^{RI}, each containing blasticidin deaminase (*bsd*) and human dihydrofolate reductase (*hdhfr*), to encode resistance to blasticidin-S and WR99210 (WR), respectively (Fig. 1a). Transfected parasites were grown on blasticidin-S to obtain 3D7/*upsC*, 3D7/*upsC*^R and 3D7/*upsC*^{RI} carrying episomes. The *hdhfr* gene represented a tool to analyse *var* promoter function, uncoupled from its chromosomal context, in its natural (WR[−]) and activated (WR⁺) state.

Growth assays showed that all three transfectants were sensitive to WR (Fig. 1b). Parasites transfected with control constructs, where the *upsC* promoter was replaced with the calmodulin (*cam*) promoter, were WR resistant. After selection of 3D7/*upsC*, 3D7/*upsC*^R and 3D7/*upsC*^{RI} with WR, resistant populations were established after six, four and ten generations, respectively. Similarly, the half-maximal inhibitory concentration (IC₅₀) for unselected populations was the same as for the WR-sensitive parent 3D7, whereas after selection they were resistant (Fig. 1c). This suggested that the *upsC*-*hdhfr* cassette was silenced and WR challenge selected for rare parasites where *upsC* was activated. To confirm this we monitored hDHFR expression by immunofluorescence (Fig. 1d). A total of 0.6% of parasites expressed hDHFR in WR-untreated parasites whereas 80% showed detectable levels after selection. When the WR-resistant population was maintained without WR for 40 generations the percentage of hDHFR-expressing cells dropped to 61%. This demonstrated that *upsC*-mediated expression was variegated and stably inherited with the same dynamics as *var* switching rates¹⁸.

In WR-untreated 3D7/*upsC*, 3D7/*upsC*^R and 3D7/*upsC*^{RI} parasites no *hdhfr* transcripts were observed, consistent with transcriptional silencing (Fig. 2a and Supplementary Fig. 2). High *hdhfr* transcript levels were present after WR selection, and the activated episomal *upsC* promoter displayed temporal transcription similar to endogenous counterparts¹⁴. Therefore the episomal promoter contained sufficient regulatory information for *upsC*-mediated control and was unaffected by the intron and/or the subtelomeric repeat element rep20 (see below). Furthermore, the silencing emanating from *upsC* had regional effects *in cis* because transcription of the *bsd* gene was substantially reduced on all silenced episomes compared to activated forms and the control plasmid pHBcam (Fig. 2a, c). Alterations in transcriptional activity were specific to *upsC* as no differences in *hdhfr* or *bsd* transcript abundance was evident in 3D7/*cam* parasites before and after WR selection (Fig. 2b). The *var* intron and rep20 played no direct role in silencing because *cam* promoter activity was unaffected by these elements (Supplementary Fig. 3). Our analysis shows that the *upsC* promoter provides sufficient information for epigenetic control to establish a silenced or activated state.

We demonstrated that the *P. falciparum* nuclear periphery was associated with transcriptional silencing⁷, and we expected that targeting of pHBupsC^R to chromosome-end clusters by inclusion of rep20 (ref. 19) would result in increased silencing. However, rep20 had no effect on *upsC*-mediated silencing (Fig. 2a). In contrast, the intron in 3D7/*upsC*^{RI} antagonized epigenetic activation and spreading of chromatin activation into the *bsd* locus (Fig. 2a). This agrees with previous results suggesting that the *upsC*-intron interaction has boundary function to protect *var* genes in chromosome-central

¹The Walter and Eliza Hall Institute of Medical Research, Parkville 3050, Australia. ²Department of Microbiology, Monash University, Clayton 3800, Australia. ³Department of Medicine, The University of Melbourne, The Royal Melbourne Hospital, Parkville 3050, Australia. ⁴Papua New Guinea Institute of Medical Research, Goroka EHP 441, Papua New Guinea.

clusters from activation through nearby euchromatic regions and neighbouring *var* genes¹⁷. In addition, the delayed resistance of 3D7/*upsC*^{RI} parasites compared to 3D7/*upsC* and 3D7/*upsC*^R parasites (Fig. 1b) indicated that activation of *upsC* paired with the intron occurred less frequently than activation of the *upsC* promoter alone. This is consistent with the role of the *var* intron in silencing^{15–17} and suggests that interplay between these elements provokes a solid and stably inherited silenced chromatin environment. However, in contrast to a recent study¹⁶ we did not find that deletion of the *var* intron was required to activate the *upsC* promoter (Supplementary Discussion).

We showed previously that gene expression at the nuclear periphery is restricted to a perinuclear region and activation and silencing of *var* genes is linked to locus repositioning⁷. We determined the nuclear position of episomal and integrated versions of active and

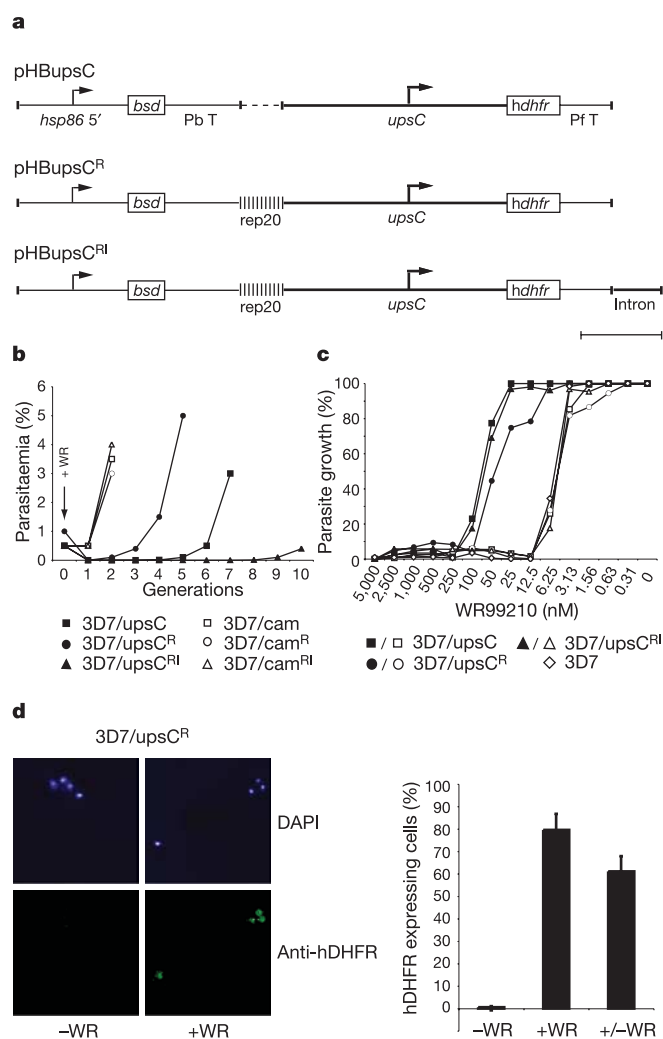


Figure 1 | WR sensitivity of *upsC* promoter transfectants. **a**, Vector maps. *hsp86* 5', heat-shock protein 86 promoter; Pb T, *P. berghei* dhfr-thymidilate synthase terminator; *upsC*, *upsC* upstream sequence¹³; Pf T, *P. falciparum* *hrp2* 3' terminator; rep20, 0.5-kb rep20 repeats; intron, 0.6-kb *var* intron sequence¹⁷. Scale bar, 1 kb. **b**, Growth assay. Parasites were challenged with WR at day 0. Assays were repeated twice with the same result. **c**, WR sensitivities before (open) and after (filled) WR selection. **d**, hDHFR expression visualized by indirect immunofluorescence assay. Percentage of hDHFR-expressing parasites before (-WR, 2 of 361) and after (+WR, 125 of 157) WR treatment, and in WR-selected parasites maintained without WR for 40 generations (+/-WR, 269 of 422). Error bars are 95% confidence intervals ($P < 0.001$). Similar results were obtained for 3D7/*upsC* and 3D7/*upsC*^{RI} (data not shown).

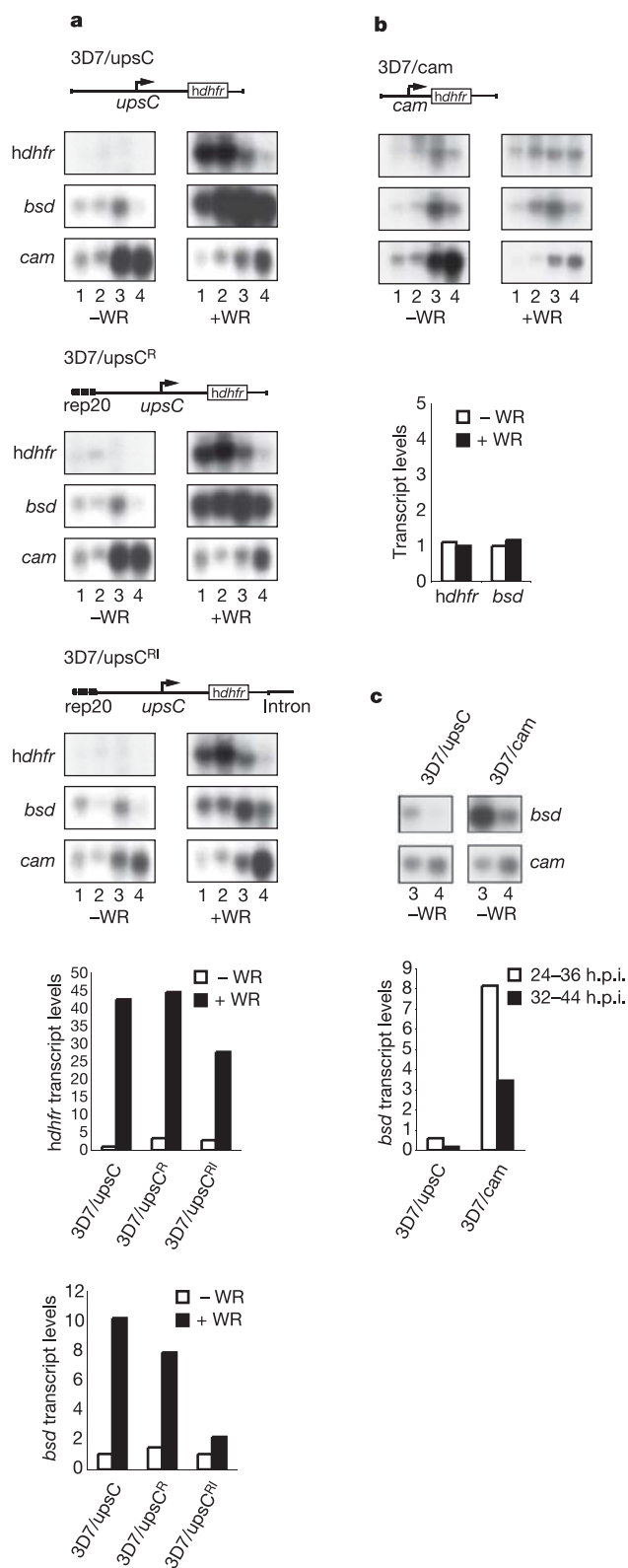


Figure 2 | Epigenetic regulation of *upsC*. **a**, Silencing and activation of *upsC*-regulated transcription. Northern blots showing episomal *hdhfr* and *bsd* transcription. Transcription of *cam* is a stage-specific loading control. Densitometry compares relative *hdhfr* and *bsd* transcript per promoter before and after selection on WR (see Supplementary Methods). **b**, *hdhfr* and *bsd* transcription in 3D7/*cam*. **c**, Spreading of *upsC*-mediated silencing. *bsd* transcription in WR-untreated 3D7/*upsC* and 3D7/*cam* parasites. The densitometry chart compares relative *bsd* transcript production per *hsp86* promoter. 1, 0–12 h post-infection (h.p.i.); 2, 12–24 h.p.i.; 3, 24–36 h.p.i.; 4, 32–44 h.p.i.

silenced *hdhfr* transgenes by fluorescent *in situ* hybridization (FISH) (Fig. 3 and Supplementary Fig. 4). In agreement with a previous report²⁰ we found silenced and active chromosome-internal *var* loci at the nuclear periphery. We also demonstrate that a chromosome-central *var* locus preferentially co-localizes with telomeric clusters, irrespective of *upsC* transcriptional state. This was not observed for episomally maintained pHBcam, which, as expected for episomes devoid of rep20 in *P. falciparum*¹⁹, showed no preferential co-localization with telomere clusters ($38 \pm 2.4\%$ co-localization). Therefore, as the parasites constitutively transcribed *bsd*, the telomeric clusters identified by co-localization with integrated pHBupsC probably represent the previously identified active perinuclear zone⁷. Surprisingly, the activated pHBupsC episome co-localized with chromosome-end clusters as frequently as integrated versions, in contrast to the silenced episomes that displayed no preferential co-localization ($P = 0.01$, paired one-tailed *t*-test) (Fig. 3b). This indicates that activation of *upsC* is restricted to a specific site associated with the active perinuclear zone. Additionally, these findings may explain why targeting of pHBupsC^R to telomeric clusters resulted in faster acquisition of WR resistance (Fig. 1b).

The above results suggested that *upsC* activation interferes with endogenous *var* transcription. To test this we analysed *var* transcription using three universal *var* probes and one probe specific to *var* PFL1960w where integration of pHBupsC occurred (see Supplementary Methods and Supplementary Fig. 4). 3D7/*upsC* parasites showed a *var* messenger RNA pattern similar to 3D7 when the plasmid-encoded *upsC* promoter was silenced (Fig. 4a). Interestingly,

PFL1960w was highly transcribed in these parasites, probably due to sequestration of this locus to the transcriptionally active perinuclear zone as a result of *bsd* transcription. After activation of *upsC* none of the probes detected any significant *var* transcription, but the *upsC*-controlled *hdhfr* was transcribed. The observation that

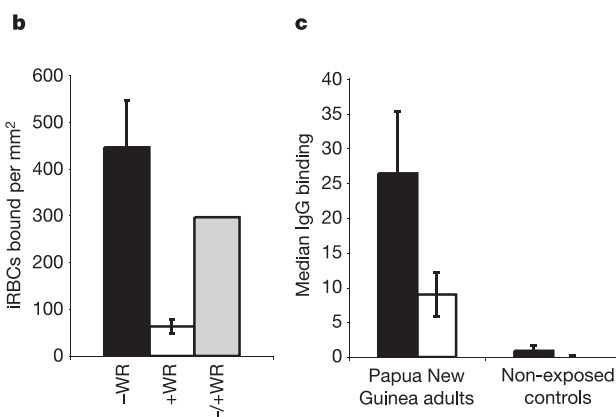
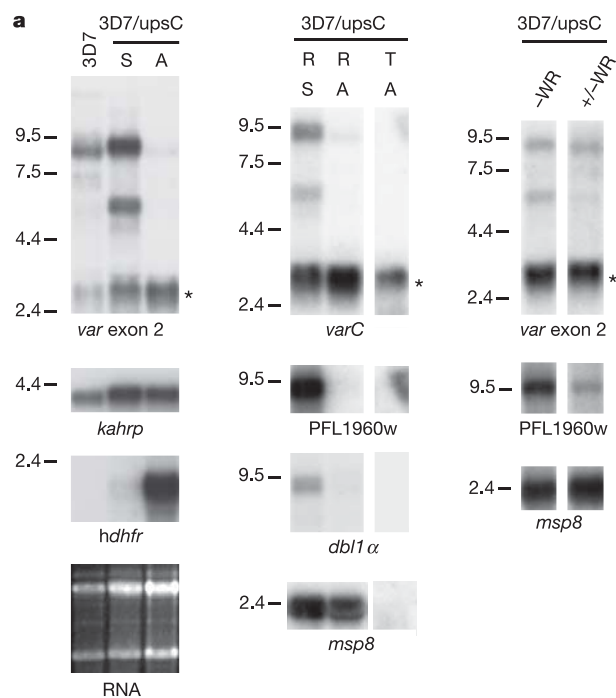


Figure 4 | Effect of *upsC* activation on mono-allelic *var* transcription.

a, Northern blot analysis. *var* transcription was monitored by hybridization with specific and universal *var* probes (Supplementary Methods). The short transcripts above 2.37 kb detected with *var* exon 2 and *varC* represent 'sterile' exon 2 transcripts³. *kahrp*, knob-associated histidine-rich protein; *msp8*, merozoite surface protein 8; S, silenced *upsC*; A, activated *upsC*; R, ring stage; T, trophozoite stage; -WR, WR-unselected parasites; -/+WR, WR-selected parasites grown in the absence of drug for 70 generations. **b**, Binding of iRBCs to CD36. Graph showing iRBC binding to CD36 for WR-selected parasites (+WR) and WR-selected parasites maintained without drug for 70 generations (-/+WR) compared to WR-unselected parasites (-WR). Values for WR-unselected and selected parasites are means ($\pm$ s.e.m.) of six experiments in duplicate with three independent transfectants (3D7/*upsC*, 3D7/*upsC*^R, 3D7/*upsC*^{R1}); values for the -/+WR parasites are means of two experiments performed in duplicate with the 3D7/*upsC* transfectant. **c**, Binding of serum IgG to iRBCs. Bar graph shows flow cytometry of IgG median binding for sera of 20 malaria-exposed adults, and five non-exposed controls, to the surface of iRBCs for WR-unselected (black) and WR-selected (white) 3D7/*upsC* parasites. Error bars are 95% confidence intervals.

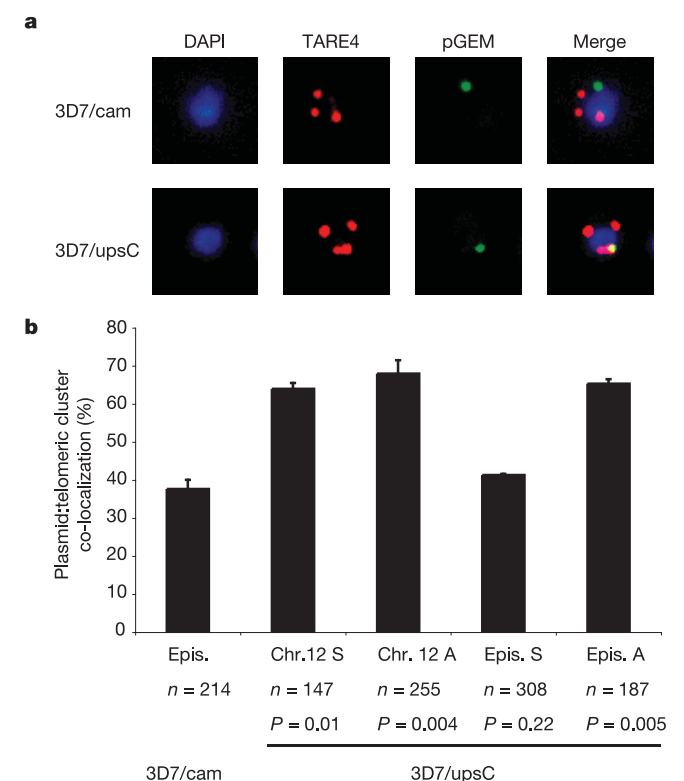


Figure 3 | Nuclear localization of silenced and activated *upsC* loci.

a, Nuclear localization of plasmids in 3D7/*upsC* and 3D7/*cam* by FISH. Epifluorescence images of nuclei (DAPI) and nuclei hybridized with TARE4 (red) and pGEM (green) to identify chromosome-end cluster and plasmid positions, respectively. **b**, Quantification of co-localization. *n*, average number of nuclei scored. Error bars are 95% confidence intervals. *P*-values were generated using a paired one-tailed *t*-test comparing the per cent co-localization in each of the 3D7/*upsC* lines to the control 3D7/*cam*. S, silenced *upsC* promoter; A, activated *upsC* promoter.

PFL1960w transcripts were absent in this population demonstrates allelic exclusion and promoter switching within a chromosome-internal *var* cluster. This knockdown of *var* transcription occurred in rings and trophozoites and was observed with 3D7/upsC^R and 3D7/upsC^{RI} carrying episomal or integrated plasmids (Fig. 4a, Supplementary Fig. 5). *var* transcripts reappeared in WR-resistant parasites after removal of WR, indicating that a subpopulation had switched from transcribing *hdhfr* to endogenous *var* genes (Fig. 4a). Therefore, *var* promoter activation is a crucial step in control of allelic *var* exclusion because transcriptional activation of one *var* promoter-associated locus inhibits transcription of the remaining *var* loci.

To confirm these observations we tested the PfEMP1-mediated binding of infected red blood cells (iRBCs) to CD36 and chondroitin sulphate A (CSA). In WR-selected parasites binding of iRBCs to CD36 was reduced to 14% of WR-untreated parasites, and WR-resistant parasites cultured in the absence of WR had intermediate binding, showing that *var* activation and PfEMP1 expression was returning as *hdhfr* was silenced (Fig. 4b). None of the parasites tested bound to CSA (not shown). In addition, IgG binding to the surface of infected erythrocytes in sera from malaria-exposed donors was significantly higher for WR-unselected compared to WR-selected 3D7/upsC parasites (median $\pm$ s.d. fluorescence 26.4 ± 20.3 versus 9.0 ± 7.2 , respectively; $P < 0.00001$, Wilcoxon's signed-rank sum test), and there was little binding of IgG among samples from non-exposed donors (median fluorescence 0.9 ± 0.9 versus 0 ± 0.3) (Fig. 4c and Supplementary Fig. 6). These findings are consistent with a substantial reduction or absence of PfEMP1 on the iRBC surface due to silencing of endogenous *var* transcription.

We have identified two important features of *var* promoters in epigenetic regulation and mono-allelic expression. First, the *upsC* *var* promoter mediates nucleation and spreading of stably inherited silenced chromatin through interactions between unidentified *cis*-acting promoter motifs and the *P. falciparum* silencing machinery. We conclude that the *upsC* promoter is required to maintain chromosome-internal *var* genes in their silenced default state, and preliminary data indicate that *upsB* promoters have a similar role in subtelomeric *var* silencing (T.S.V. *et al.*, unpublished data). Second, the information provided by the *upsC* promoter is sufficient to insert a locus into the allelic exclusion programme of *var* transcription independent of chromosomal context (Supplementary Discussion). On the basis of our findings, the simplest model for mechanisms underlying mono-allelic *var* expression is a unique perinuclear compartment, associated with the active chromosome-end cluster, responsible for transcription of a single *var* locus (Supplementary Fig. 1). Switching in *var* transcription, and thus antigenic variation of PfEMP1, would occur by competition of silenced *var* promoters for occupancy of this compartment. A similar mechanism has been identified to control mono-allelic expression of *vsg* genes in African trypanosomes²¹, and has been proposed to control mono-allelic transcription of the mammalian odorant receptor gene family²². The results presented here are, to our knowledge, the first description of transcriptional repression of a gene family and are fundamental to understanding antigenic variation and epigenetic regulation in *P. falciparum* and mono-allelic transcription and phenotypic variation in other systems. The *var* knockdown parasites obtained in this study provide an invaluable tool to understand the role of PfEMP1 in parasite-mediated virulence, protective immunity and cyto-adherence.

METHODS

Parasites, drug sensitivity and cyto-adherence. *P. falciparum* 3D7 and transfected parasites were cultured as described previously²³. Growth synchronization was achieved using sorbitol. A total of 250 μ l of packed iRBCs (5–10% ring parasites) were transfected by electroporation of 100 μ g of plasmid. Selection on blasticidin-S (2 μ g ml⁻¹) was 4–6 h after transfection. Selection on WR99210

was at 4 nM. Drug sensitivity assays were performed with standard techniques. Cyto-adherence of parasitized erythrocytes to CD36 and CSA were as described²⁴.

Transfection constructs. The puromycin resistance cassette in pHH1/pac²⁵ was excised with *NotI*/BglII. After ligation of adaptors the cassette was cloned into *SacII*/*NotI*-digested pHHMC*/R0.5 (ref. 19) to generate pHH_VP, where a 0.5-kilobase (kb) rep20 fragment separates the head-to-tail expression cassettes. The vector pHBcam^R was obtained by replacing the puromycin resistance gene with *bsd* amplified from pCBM²⁶. pHBcam^{RI} was generated by ligation of a 0.6-kb *var* intron fragment¹⁷ into the *EcoRI* site of pHBcam^R. pHBupsC^R was generated by replacing the *BglII*/*BamHI* *cam* promoter with the *upsC* promoter (PFL1960w) amplified from pCAT5B1 (ref. 13). pHBupsC^{RI} was derived in the same way from pHBcam^{RI}. The 0.5-kb rep20 sequence in pHBcam^R and pHBupsC^R was deleted by digestion with *PstI*/BglII, end-polishing and re-ligation to yield pHBcam and pHBupsC, respectively.

Southern and northern analysis. gDNA was digested with *EcoRV* and *PvuII*. Southern blots were probed with *hdhfr* and *psfr2* fragments. Copy number was determined by densitometry comparing signal intensities of *EcoRV*/*PvuII* fragments representing *hdhfr* to the *EcoRV* fragment from the single copy *psfr2*. Plasmid integration was mapped by hybridization with *hdhfr* and a 735-base pair fragment derived from the 5' end of PFL1960w. Pulsed-field gel electrophoresis (PFGE) was carried out as described²⁷. Northern experiments are described in detail in Supplementary Methods.

Indirect immunofluorescence assay and FISH. Indirect immunofluorescence assays to detect hDHFR expression were performed as described²⁸ using a monoclonal mouse anti-bovine dihydrofolate reductase antibody (BD Biosciences) and anti-mouse Alexa-Fluor 488-nm (Molecular Probes). Parasite DNA was visualized using DAPI and cells were counted for DAPI and hDHFR positivity. FISH analysis was performed as presented elsewhere⁷ using telomere-associated repeat element 4 (TARE4) and pGEM probes for the detection of telomere clusters and the plasmid backbone, respectively. Images were counted by three independent scorers in a blinded design in two to three independent FISH experiments.

Fluorescence-activated cell sorting. Serum samples were tested for IgG bound to the surface of mature trophozoite-infected erythrocytes using flow cytometry²⁹. Cells were sequentially incubated with test serum or plasma diluted 1/10, rabbit anti-human IgG (Fc-specific, Dako; 1:100), Alexa-Fluor 488-conjugated anti-rabbit Ig (Molecular Probes; 1:750) and ethidium bromide (10 μ g ml⁻¹). Samples were analysed using a FACSCalibur flow cytometer (Becton-Dickinson) and Flowjo software (TreeStar). For each sample the mean fluorescence of uninfected erythrocytes was deducted from the mean fluorescence of infected erythrocytes. All samples were tested in duplicate. Serum samples were obtained from malaria-exposed adults from Madang, Papua New Guinea (ten men, five non-pregnant women, five pregnant women) following informed consent. Five samples from non-exposed Australian residents were included as controls. Ethical clearance was obtained from the Medical Research Advisory Committee, Department of Health, Papua New Guinea.

Received 1 September; accepted 4 November 2005.

Published online 28 December 2005.

- Biggs, B. A. *et al.* Antigenic variation in *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **88**, 9171–9174 (1991).
- Roberts, D. J. *et al.* Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature* **357**, 689–692 (1992).
- Su, X. Z. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**, 89–100 (1995).
- Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum* *var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101–110 (1995).
- Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77–87 (1995).
- Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
- Duraing, M. T. *et al.* Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell* **121**, 13–24 (2005).
- Freitas-Junior, L. H. *et al.* Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* **121**, 25–36 (2005).
- Chen, Q. *et al.* Developmental selection of *var* gene expression in *Plasmodium falciparum*. *Nature* **394**, 392–395 (1998).
- Scherf, A. *et al.* Antigenic variation in malaria: *in situ* switching, relaxed and mutually exclusive transcription of *var* genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.* **17**, 5418–5426 (1998).

11. Jensen, A. T. *et al.* *Plasmodium falciparum* associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes. *J. Exp. Med.* **199**, 1179–1190 (2004).
12. Deitsch, K. W., del Pinal, A. & Wellems, T. E. Intra-cluster recombination and var transcription switches in the antigenic variation of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **101**, 107–116 (1999).
13. Voss, T. S. *et al.* Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum* var gene 5' flanking sequences. *Mol. Biochem. Parasitol.* **107**, 103–115 (2000).
14. Voss, T. S., Kaestli, M., Vogel, D., Bopp, S. & Beck, H. P. Identification of nuclear proteins that interact differentially with *Plasmodium falciparum* var gene promoters. *Mol. Microbiol.* **48**, 1593–1607 (2003).
15. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in var genes. *Nature* **412**, 875–876 (2001).
16. Gannoun-Zaki, L., Jost, A., Mu, J., Deitsch, K. W. & Wellems, T. E. A silenced *Plasmodium falciparum* var promoter can be activated *in vivo* through spontaneous deletion of a silencing element in the intron. *Eukaryot. Cell* **4**, 490–492 (2005).
17. Calderwood, M. S., Gannoun-Zaki, L., Wellems, T. E. & Deitsch, K. W. *Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. *J. Biol. Chem.* **278**, 34125–34132 (2003).
18. Horrocks, P., Pinches, R., Christodoulou, Z., Kyes, S. A. & Newbold, C. I. Variable var transition rates underlie antigenic variation in malaria. *Proc. Natl Acad. Sci. USA* **101**, 11129–11134 (2004).
19. O'Donnell, R. A. *et al.* A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of *Plasmodium falciparum* chromosomes. *EMBO J.* **21**, 1231–1239 (2002).
20. Ralph, S. A., Scheidig-Benatar, C. & Scherf, A. Antigenic variation in *Plasmodium falciparum* is associated with movement of var loci between subnuclear locations. *Proc. Natl Acad. Sci. USA* **102**, 5414–5419 (2005).
21. Navarro, M. & Gull, K. A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. *Nature* **414**, 759–763 (2001).
22. Shykind, B. M. Regulation of odorant receptors: one allele at a time. *Hum. Mol. Genet.* **14** (suppl. 1), R33–R39 (2005).
23. Trager, W. & Jensen, J. B. Cultivation of malarial parasites. *Nature* **273**, 621–622 (1978).
24. Beeson, J. G. *et al.* *Plasmodium falciparum* isolates from infected pregnant women and children are associated with distinct adhesive and antigenic properties. *J. Infect. Dis.* **180**, 464–472 (1999).
25. de Koning-Ward, T. F., Waters, A. P. & Crabb, B. S. Puromycin-N-acetyltransferase as a selectable marker for use in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **117**, 155–160 (2001).
26. Mamoun, C. B., Gluzman, I. Y., Goyard, S., Beverley, S. M. & Goldberg, D. E. A set of independent selectable markers for transfection of the human malaria parasite *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96**, 8716–8720 (1999).
27. Rubio, J. P., Thompson, J. K. & Cowman, A. F. The var genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. *EMBO J.* **15**, 4069–4077 (1996).
28. Tonkin, C. J. *et al.* Localization of organellar proteins in *Plasmodium falciparum* using a novel set of transfection vectors and a new immunofluorescence fixation method. *Mol. Biochem. Parasitol.* **137**, 13–21 (2004).
29. Beeson, J. G. *et al.* Antibodies to variant surface antigens of *Plasmodium falciparum*-infected erythrocytes and adhesion inhibitory antibodies are associated with placental malaria and have overlapping and distinct targets. *J. Infect. Dis.* **189**, 540–551 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements We are grateful to A. Gemmill and J. Baum for statistical advice. This work was supported by the National Health and Medical Research Council and the Wellcome Trust. A.F.C. and B.S.C. are International Fellows of the Howard Hughes Medical Institute. T.S.V. is supported by fellowships from the Swiss National Science Foundation and the Roche Research Foundation. A.J.M. is supported by an Australian Postgraduate Award. J.G.B. was supported by the NHMRC and the Miller Fellowship of WEHI. We also thank G. Kelly and A. Raiko for sample collection and processing in Papua New Guinea.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.F.C. (cowman@wehi.edu.au).

Identification of pathways regulating cell size and cell-cycle progression by RNAi

Mikael Björklund^{1,3*}, Minna Taipale^{1,3*}, Markku Varjosalo^{1,3}, Juha Saharinen^{3,4}, Juhani Lahdenperä² & Jussi Taipale^{1,2,3}

Many high-throughput loss-of-function analyses of the eukaryotic cell cycle have relied on the unicellular yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. In multicellular organisms, however, additional control mechanisms regulate the cell cycle to specify the size of the organism and its constituent organs¹. To identify such genes, here we analysed the effect of the loss of function of 70% of *Drosophila* genes (including 90% of genes conserved in human) on cell-cycle progression of S2 cells using flow cytometry. To address redundancy, we also targeted genes involved in protein phosphorylation simultaneously with their homologues. We identify genes that control cell size, cytokinesis, cell death and/or apoptosis, and the G1 and G2/M phases of the cell cycle. Classification of the genes into pathways by unsupervised hierarchical clustering on the basis of these phenotypes shows that, in addition to classical regulatory mechanisms such as Myc/Max, Cyclin/Cdk and E2F, cell-cycle progression in S2 cells is controlled by vesicular and nuclear transport proteins, COP9 signalosome activity and four extracellular-signal-regulated pathways (Wnt, p38 β MAPK, FRAP/TOR and JAK/STAT). In addition, by simultaneously analysing several phenotypes, we identify a translational regulator, eIF-3p66, that specifically affects the Cyclin/Cdk pathway activity.

The cell cycle can be divided into distinct phases including a synthesis (S) phase, where DNA is replicated, and a mitosis (M) phase, where cell division occurs. In animal cells, growth and the synthesis of components required for these phases are regulated by extracellular growth factors and occur mainly in two gap phases, G1 (between M and S) and G2 (between S and M)^{2–4}. The ordered progression of the cell-cycle phases is orchestrated by cyclin-dependent kinases (Cdks), whose activity is controlled by phosphorylation and by association with specific regulatory subunits, the cyclins^{2–4}.

The cell cycle is a robust system in which compensatory mechanisms control its overall length⁵. We therefore analysed the effect of RNA-mediated interference (RNAi)-induced loss of function of *Drosophila* genes in S2 cell-cycle distribution using flow cytometry, which allows direct determination of the fraction of cells in different phases of the cell cycle (Supplementary Fig. S1).

Consistent with the high efficiency of RNAi in *Drosophila*^{6–9}, clear phenotypes including arrest in G1, G2/M and S, cytokinesis and DNA replication defects, and apoptosis and/or cell death were observed when targeting known regulators of these processes (Fig. 1a and Supplementary Fig. S2). We next carried out a pilot screen targeting *Drosophila* kinases and phosphatases individually (Fig. 1b and Supplementary Fig. S3), and in pools of homologues to address redundancy (Fig. 1c and Supplementary Fig. S4). Unexpectedly, the analysis of redundancy did not reveal any additional cell-cycle regulators (Fig. 1c and Supplementary Table S3). These results

suggest that a considerably lower fraction of *Drosophila* kinases regulate the cell cycle (19%) than has been reported previously (35%, ref. 10; for explanation and comparison, see Supplementary Table S4 and Fig. S5). Despite this, our screen identified also kinases for which a role in the S2 cell cycle has not been appreciated (for example, AKT1, CG7177 and MEKK1/MEKK4).

We reasoned that, by screening most *Drosophila* genes, we should be able to identify the pathways regulating the S2 cell cycle. Screening of 11,971 double-stranded RNAs (dsRNAs) generated from the *Drosophila* Gene Collection (DGC) releases 1 and 2 (see Supplementary Methods and refs 7, 8) showed that loss of 270 and 169 genes resulted in significant changes in G1 and G2 populations, respectively (Supplementary Fig. S6). A strong correlation between cell size at the G1 and G2 phases was observed in individual samples, suggesting that *Drosophila* cells do not have a 'strong' cell-size checkpoint¹¹ that forces cells to a particular size at a defined cell-cycle phase (Supplementary Fig. S7). Components acting on the same direction in a particular process, such as ribosomal proteins or Dp, E2f and Cyclin E, appeared in discrete areas of a plot describing G1 cell size as a function of fraction of cells in G1 (Fig. 2a). The strongest phenotypes, characterized by specifically decreased cell size, were observed with dsRNAs targeting *Rbf*, a negative regulator of E2f pathway, and *cropped* (*crp*), the *Drosophila* AP4 transcription factor (Supplementary Tables S3 and S5).

To identify genes involved in cell death and cytokinesis, we analysed changes in the populations of cells with less than 2N DNA or more than 4N DNA, respectively (Fig. 2 and Supplementary Fig. S8). The increase in the <2N population in all cases consisted of apoptotic or dead cells; the genes whose loss caused this phenotype included previously unknown effectors, the known inhibitor of apoptosis *Thread* and several mitosis regulators, consistent with induction of apoptosis by mitotic catastrophe¹² (Fig. 2b, and Supplementary Figs S8 and S9 and Table S6). The increase in the >4N population was due to cells that had undergone two rounds of DNA replication without cytokinesis (8N DNA); dsRNAs causing this phenotype targeted several previously undescribed genes and known cytokinesis regulators⁹ (Fig. 2c, and Supplementary Fig. S8 and Table S6).

There have been several large-scale RNAi screens, and in many there is very little overlap between the gene sets identified even when the same phenotypes are analysed. This is apparently due to a very large number of false positives (see Supplementary Tables S7 and S8, and Fig. S10). To reduce the number of false positives, we analysed all phenotypes by using criteria that excluded all 564 control samples (>99.99994% confidence). Roughly 4% of genes (488; Supplementary Table S3) had a cell-size, cell-cycle or cell-death phenotype. From an analysis of 19 different pathways and/or protein complexes, we estimate that our screen identified ~80% of strong non-redundant

¹Molecular and Cancer Biology Program, and ²High Throughput Center, Biomedicum Helsinki, PO Box 63 (Haartmaninkatu 8), FI-00014 University of Helsinki, Finland.

³Department of Molecular Medicine, National Public Health Institute, and ⁴Biomedicum Bioinformatics Unit, Biomedicum, FI-00251 Helsinki, Finland.

*These authors contributed equally to this work.

cell-cycle regulators (Fig. 3a and Supplementary Table S9). Despite the general effectiveness of the RNAi (Fig. 3a and Supplementary Fig. S11), however, we did not identify some genes with clearly defined functions in the cell cycle (Fig. 3a and Supplementary Fig. S12). In different cases, the weak RNAi phenotype could be explained by high protein levels or stability, redundant function and/or the fact that defects in mitotic or DNA replication fidelity often do not appreciably alter overall cell-cycle phase distribution¹³ (Supplementary Table S9 and Fig. S12).

Of the 78 known interactions between the genes identified, 70 were consistent with the observed phenotypes (Supplementary Table S10). A high proportion of the diseases linked to the human orthologues of the identified genes are related to cancer (8 of 25; Supplementary Table S5), consistent with the known relationship between cell-cycle regulation and cancer. As expected, the *Drosophila* protein–protein interaction map¹⁴ indicated a much larger number of interactions between proteins than we identified than between a random set of proteins (Fig. 3b), and most interactions were consistent with the observed phenotypes (107 of 143; Supplementary Table S11).

Of the 488 genes that we identify, 319 (65%; Supplementary Tables S1 and S3) have not been identified in previous *Drosophila* and mammalian RNAi screens. A principal difference between our screen and previous studies^{6,9} is that, by using flow cytometry, we can simultaneously identify six distinct phenotypes, allowing unbiased classification of the identified genes into pathways using hierarchical clustering (Fig. 3c and Supplementary Fig. S13). Positive components of a given pathway should segregate into the same cluster, because their loss is expected to cause similar phenotypes. This was indeed observed, for example 43 of 56 dsRNAs in one 'translation' cluster targeted ribosomal subunits. Another 'G1 cell-cycle' cluster contained the classical positive G1 regulators of the Cdk/E2f pathway, including E2f, Dp, Cyclin E, Cdk2 and Cdk4 (Fig. 3c). Because of an opposite phenotype, Rbf, a known negative regulator of this pathway, was not included in this cluster.

The other genes in the G1 cell-cycle cluster included two cullins (Cul-1/lin19 and Cul-4), involved in ubiquitin-mediated protein degradation, and five components of the COP9 signalosome (Fig. 3c), a protein complex linked to regulation of cullin activity¹⁵. In

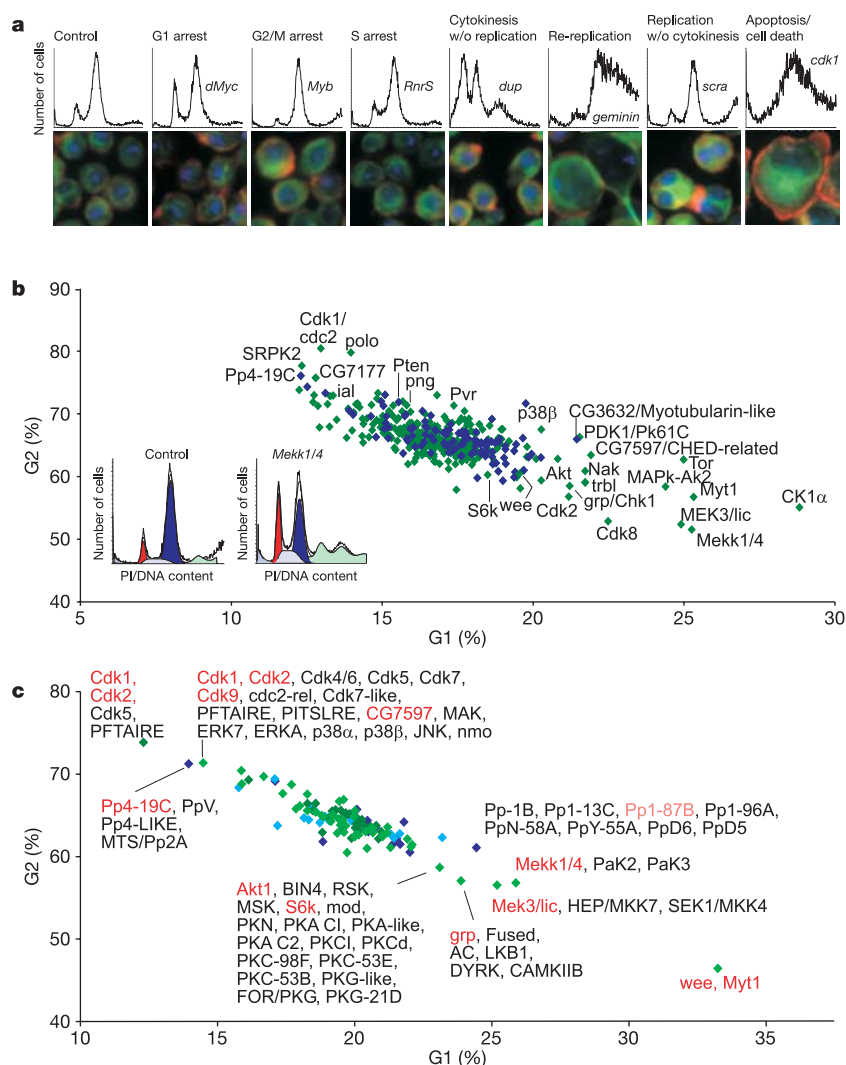


Figure 1 | Cell-cycle regulation by protein phosphorylation in S2 cells.

a, Flow cytometry can detect multiple RNAi phenotypes. Shown are histograms (top; DNA content) and micrographs (bottom; blue, DNA; red, F-actin; green, α -tubulin) of dsRNA-treated cells. Note that, similar to *geminin* dsRNA, *Cdk1* dsRNA causes cell death and also a giant cell and re-replication phenotype. *scra* dsRNA induces binucleate cells with 8N DNA, whereas *Myb* dsRNA induces a small increase in cells having single nuclei with 8N DNA. **b**, **c**, Redundancy of protein kinase (green) and phosphatase

(blue) function. Fraction of G1 (x axis) and G2 (y axis) phase cells after 4 d of treatment with individual dsRNAs (**b**) or pooled dsRNAs (**c**) against related genes (dark colour, close homologues; light colour, more distant homologues; red text, genes identified previously in individual screen). Note that a pool containing *wee* and *Myt1* has a much stronger phenotype than *wee* or *Myt1* alone. *Pp1-87B* (pink) was identified only in the DGC screen. Inset, phenotypic analysis of the control and MEKK1/MEKK4 samples.

mammals Cul-4 has been linked to DNA damage response¹⁶, but our results indicate that Cul-4 is also involved in normal cell-cycle regulation in animals. Consistent with the similar phenotypes of the cullins, COP9 subunits (CSNs) and several Cdk/E2f pathway components (Fig. 3c), the increase in G1 induced by dsRNAs targeting CSNs could be reversed by simultaneously targeting the Cdk inhibitor *dacapo* (Supplementary Fig. S15). Our results are thus consistent with a loss of COP9 leading to G1 arrest owing to failure of cullin-mediated degradation of *dacapo*. The positive role of COP9 in the cell cycle is in agreement with mouse knockout studies^{15,17}. In *Drosophila* ovaries, however, COP9 negatively regulates the cell cycle¹⁸, suggesting that its effects may be tissue-specific.

Notably, a translational initiation factor, *eIF-3p66*, clustered with the G1 cell-cycle regulators. Expression of *eIF-3p66* also correlates well with that of Rbf, E2f and Cyclin E during *Drosophila* development (Supplementary Fig. S14). dsRNAs targeting *eIF-3p66* resulted in an increased cell size, indicating that, unlike loss of other translational regulators, loss of *eIF-3p66* does not generally inhibit translation. Furthermore, the phenotype caused by dsRNA targeting *eIF-3p66* was reversed by dsRNA targeting *dacapo* (Supplementary Fig. S15 and data not shown), indicating that *eIF-3p66* specifically affects the Cyclin/Cdk pathway.

The cluster of genes whose loss resulted in an increase in G1 cell size with no G1 DNA content phenotype contained several genes involved in transcription and messenger RNA processing (Fig. 3c). The lack of G1 phenotype is probably due to a uniform effect of transcriptional inhibition on all cell-cycle phases, and continued growth of the cells is possible owing to translation of stable mRNAs. The distinct phenotype of transcriptional and translational regulators suggests that cell size may be controlled at the RNA level, either

directly by regulatory RNAs or through differential stability of mRNAs of proteins controlling cell division and growth.

A number of genes whose loss resulted in increased G1 content were classified into a defined 'signalling' cluster, which contained two regulators of translation (*eIF-4G* and *E1f1α48D*), a ubiquitin E3-ligase (*URE-B1/CG8184*), and components of several extracellular ligand-regulated signalling pathways (Fig. 3c). These genes differed from those in the G1 cell-cycle cluster in that the dsRNAs targeting them resulted in decreased cell size. Thus, the signalling pathways either affect cell division through regulation of growth, or affect both cell growth and cell division in parallel (see ref. 19).

By manually comparing the genes identified (Supplementary Table S3) with literature, we identified four signalling pathways that regulate the cell cycle: the FRAP/TOR, JAK/STAT, Wnt and p38βMAPK pathways. The FRAP/TOR²⁰ pathway is known to regulate cell proliferation and, consistently, we identified eight known components of this pathway as G1 regulators (Supplementary Fig. S15). In S2 cells, another known cell-cycle regulatory pathway, the JAK/STAT pathway²¹, had a relatively weak effect, as indicated by the mild phenotypes observed by dsRNAs targeting several of its components (Supplementary Fig. S15). The identified genes known to regulate the Wnt pathway (Supplementary Fig. S15) indicated that this pathway is not active in S2 cells, but that treatments that activate it inhibit S2 cell proliferation.

Several genes linked to the p38βMAPK pathway^{22,23} (Fig. 1b and Supplementary Table S1) were also identified, and we verified the function of MEKK1/MEKK4 and MEK3 in this pathway by analysing p38βMAPK phosphorylation in response to dsRNA treatments (Supplementary Fig. S15). Although regulation of the cell cycle by the p38βMAPK pathway in the absence of stress or DNA damage has

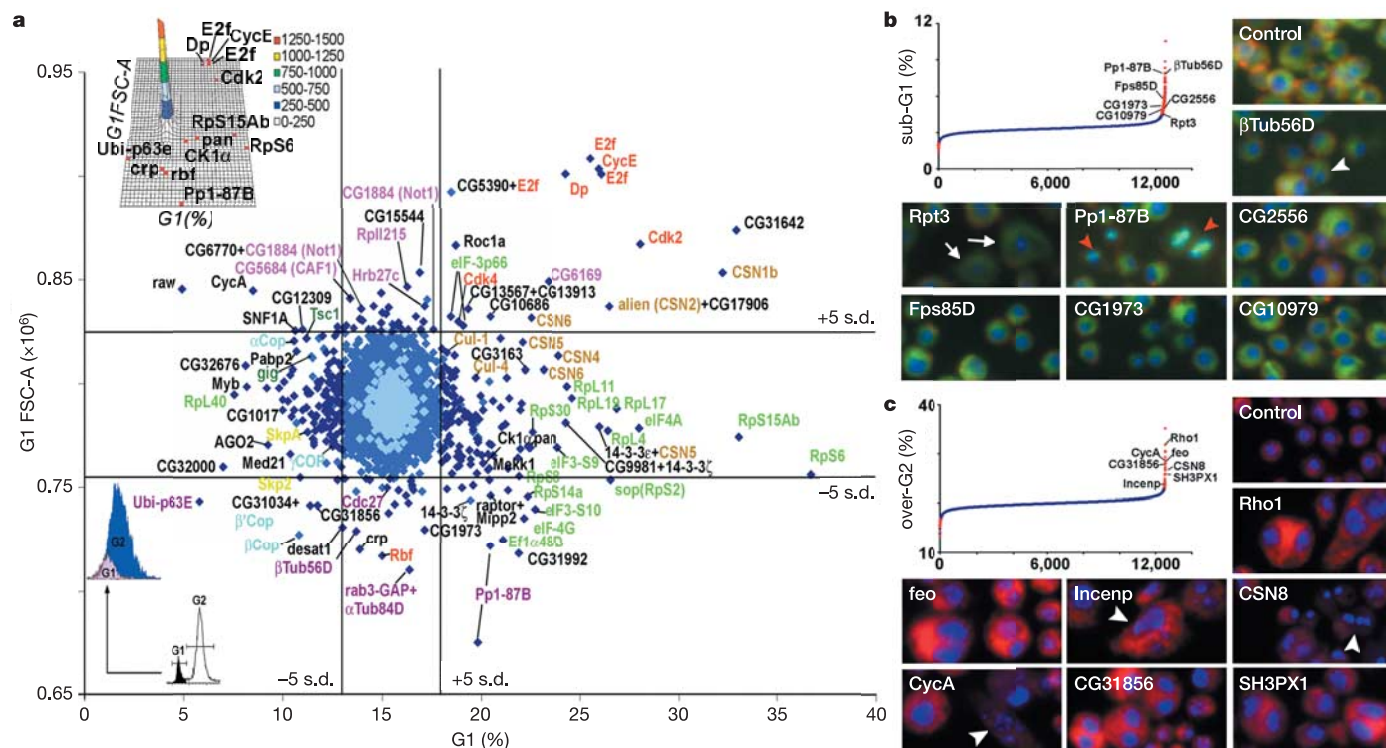


Figure 2 | The DGC RNAi library screen. **a**, Cell size (FSC-A) as a function of fraction of cells in G1. Control samples are indicated in light blue, and samples that do not differ significantly from the mean (Supplementary Table S3) in blue and dark blue, respectively. Genes involved in the Cdk/E2f pathway (red), COP9 signalosome activity (brown), translation (light green), apoptosis and/or cell death (purple; see Supplementary Fig. S6a), vesicular transport (cyan), SCF^{Skp2} (yellow-green), RNA transcription or splicing (light purple), and tuberculosis

(dark green) are indicated. Lines indicate 5 s.d. of control samples. Inset, sample density (z axis) as a function of G1 DNA content (x axis) and G1 size (y axis). **b**, dsRNAs inducing cell death (blue, DNA; red, F-actin; green, α-tubulin). Note that some dsRNAs also induce a mitotic (red arrowhead), flat cell (arrow) or multinucleate (white arrowhead) phenotype. **c**, dsRNAs inducing over-4N DNA content (blue, DNA; red, protein). Arrowheads indicate abnormal nuclei. Graphs in **b** and **c** show ranking of phenotypic scores; genes outside the 5 s.d. limits are in red.

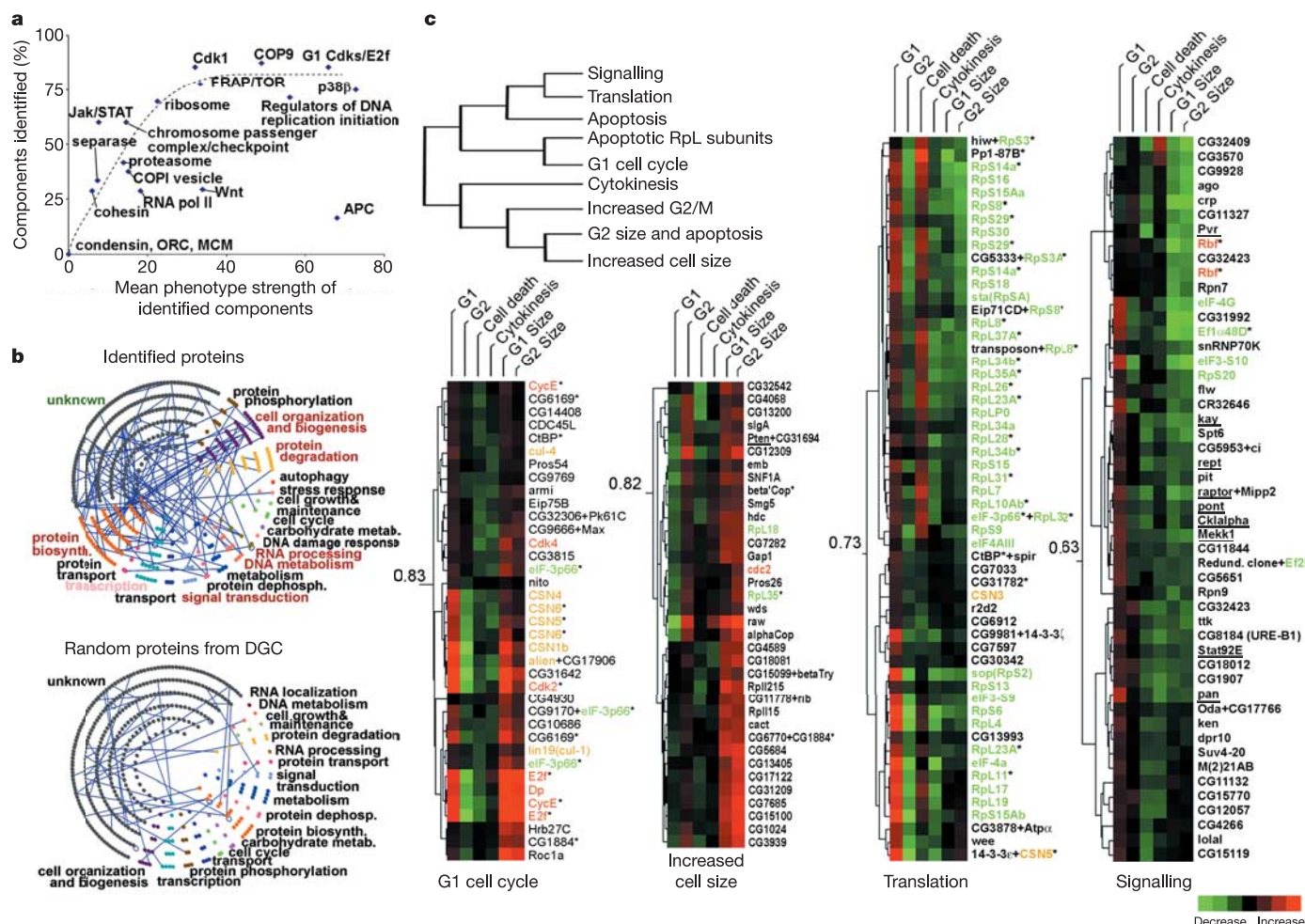


Figure 3 | Classification of genes identified in the DGC screen.

a, Percentage of genes identified in the 19 indicated pathways or protein complexes increases as a function of the mean phenotype strength of the complex (approximated by the dotted line). **b**, Identified proteins (top) have a much larger number of interactions¹⁴ (coloured lines) than does a random set of proteins (bottom; the median example of five random sets is shown). Overrepresented (red, $>2\times$; light red, $>1.5\times$) and underrepresented (green) gene ontology classes are indicated. **c**, Hierarchical clustering of

genes based on their phenotypic scores. Top left, major sub-branches, named according to phenotype and/or known activity of the included genes. Four sub-branches are shown in detail (for complete analysis, see Supplementary Fig. S13). Note that genes that are present more than once (asterisks) typically cluster tightly and that genes involved in the Cdk/E2f pathway (red), COP9 signalosome activity (yellow), translation (light green) and signal transduction (underlined) are predominantly found in specific clusters. Pearson correlation coefficients of the clusters are also indicated.

not previously been appreciated, our data are consistent with the identification of MAPK-Ak2 as a checkpoint kinase²⁴.

Loss of several kinases that negatively regulate Cdk1 activity led to an increase in G1 content (Supplementary Fig. S15), probably owing to an accelerated G2/M phase⁵. We also observed that Chk1 and MAPK-Ak2, two kinases that have been linked to DNA damage checkpoints^{24–26}, operate in normal cell-cycle regulation of cultured S2 cells. Together with the finding that in human cells Chk1 shields Cdk1 from premature activation²⁷, our results suggest that during normal cell cycle these kinases have basal activity that slows down G2 progression. The basal activity could represent partial activation that occurs locally in response to damage during normal replication, possibly effecting localized changes in DNA replication and repair.

In summary, we have reported that several pathways and processes are involved in cell-cycle and cell-size regulation, including basal transcription, vesicular trafficking, nuclear export, ubiquitin-mediated protein degradation, and the Cdk/E2f, JAK/STAT, FRAP/TOR, p38 β MAPK and Wnt signalling pathways (Supplementary Table S1). The genes identified are likely to include previously unknown components of these pathways (Fig. 3a), ubiquitin ligase and kinase substrates (Supplementary Table S12), and targets of transcription factors (Supplementary Table S13) such as E2f and

Myc/Max that are relevant for cell-cycle progression. Thus, our results will also serve as a solid foundation for a systems biology analysis of the metazoan cell cycle.

METHODS

RNAi libraries, cell culture and transfections. The RNAi libraries targeting *Drosophila* protein and lipid kinases and phosphatases, and the genes in the publicly available clone collections (BACPAC Resources Centre), consisting of about 70% of the genes annotated by the Berkeley *Drosophila* genome project^{28,29}, are described in the Supplementary Methods (for sequence information, see <http://www.fruitfly.org/sequence/dlcDNA.shtml>). Each dsRNA was screened individually. The DGC libraries occasionally contain clones in which sequences from two genes were ligated during cDNA cloning. In such clones, both genes are indicated separated by a plus sign, as the dsRNA transcribed from such a clone will target both genes. Single gene targeting could identify some genes with overlapping function (see Fig. 1). Although simultaneous targeting is required for identifying redundant genes, identification of genes when targeted alone does not rule out overlapping functions. If multiple proteins have similar activity that is rate limiting for cell-cycle progression, loss of even one protein should result in a weak phenotype. Because most pools containing known effectors were identified (Fig. 1), pooling did not substantially reduce the RNAi efficiency.

Drosophila S2 cells were cultured at controlled temperature (23.5 °C; Binder

KB cooling incubator) in *Drosophila* SFM (Invitrogen) supplemented with 10% fetal bovine serum and antibiotics. Under these culture conditions, most of the S2 cells were in the G2 phase of the cell cycle with 4N DNA content, as described previously³⁰. For DGC screen transfections, 80,000 S2 cells were plated on all wells of 96-well plates. On the following day, the cells were transfected with dsRNAs in triplicate using Effectene (Qiagen) according to manufacturer's instructions, except that 50 ng dsRNA, 0.4 µl enhancer and 0.5 µl Effectene in 50 µl EC buffer were used per well (25 ng dsRNA, 25% transfection reagents and 10,000 cells were used in 384-well plates). Transfections were done similarly for the kinase/phosphatase screen, except that 450,000 cells, 125 ng dsRNA, 0.8 µl enhancer and 5 µl Effectene were used in 24-well plates (Fig. 1 and Supplementary Fig. S3).

Flow cytometry. Cells cultured for 4 days after transfection were collected by centrifugation, the culture media was removed, and the cells were fixed with ice-cold 70% ethanol overnight. The cells were then suspended in PBS containing 30 µg ml⁻¹ propidium iodide (P1304MP; Molecular Probes) and 30 µg ml⁻¹ RNase A (R5125; Sigma). After 30 min incubation at 37 °C, the cells were analysed with a FACSArray (DGC screen) or LSR (kinase/phosphatase screen) flow cytometer (Beckton Dickinson). Cells that were not immediately analysed were stored at 4 °C, and equilibrated to room temperature before analysis. The dsRNAs were screened individually in triplicate, and phenotypes were scored by automated analysis. A minimum confidence interval of 99.99994%, excluding all 564 'mock transfection' control samples, was used to identify hits; thus, the data set is essentially free of statistical error, as only 0.04 false positives should occur in the 72,000 hypotheses (12,000 × 6 phenotypes) tested. Data analysis is described in detail in the Supplementary Methods. Small fluctuations in G1 and G2 peak widths prevented statistically reliable automated analysis of S phase. Visual inspection was additionally used to classify sub-G1 and over-G2 phenotypes. Apart from *double parked*, *Cdk1* and *geminin*, which were all re-identified in our screen, no other dsRNAs resulted in a division without replication or (aneuploid) DNA re-replication phenotypes (see Fig. 1a). Mean phenotype strength of complexes was calculated by using the strongest of the six measured phenotypes (normalized to 0–100%; see Supplementary Table S9). Only genes that scored as hits were included in the analysis to determine the hit rate as a function of mean phenotype strength.

MTS assay and microscopy. MTS metabolic activity assay was assayed using a CellTiter96AQ kit (Promega) according to the manufacturer's instructions in triplicate 384-well plates after 4 days of dsRNA treatments. For microscopy, cells were plated on 96- or 384-well Packard ViewPlates, transfected with dsRNAs, and stained after 4 days with Hoechst33342 (Invitrogen) and Celltrace Far-Red SSAOE whole protein stain (for counting of multinucleate cells; Invitrogen) or with Hoechst33342, AlexaFluor546-conjugated phalloidin (Invitrogen) and fluorescein isothiocyanate (FITC)-conjugated anti-α-tubulin (for analysis of mitosis). The cells were imaged and analysed by a Celloomics ArrayScan high content microscope with the Morphology Explorer BioApplication. Total cell counts were obtained using automated analysis, and the mitotic and multinucleated cells were counted visually by a researcher who was blind to the samples. Antibodies are described in Supplementary Methods.

Received 13 June; accepted 21 November 2005.

- Conlon, I. & Raff, M. Size control in animal development. *Cell* **96**, 235–244 (1999).
- Koepp, D. M., Harper, J. W. & Elledge, S. J. How the cyclin became a cyclin: regulated proteolysis in the cell cycle. *Cell* **97**, 431–434 (1999).
- Murray, A. W. Recycling the cell cycle: cyclins revisited. *Cell* **116**, 221–234 (2004).
- Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* **432**, 316–323 (2004).
- Reis, T. & Edgar, B. A. Negative regulation of dE2F1 by cyclin-dependent kinases controls cell cycle timing. *Cell* **117**, 253–264 (2004).
- Boutros, M. *et al.* Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835 (2004).
- Lum, L. *et al.* Identification of Hedgehog pathway components by RNAi in *Drosophila* cultured cells. *Science* **299**, 2039–2045 (2003).
- Kim, Y. O., Park, S. J., Balaban, R. S., Nirenberg, M. & Kim, Y. A functional genomic screen for cardiogenic genes using RNA interference in developing *Drosophila* embryos. *Proc. Natl Acad. Sci. USA* **101**, 159–164 (2004).
- Eggert, U. S. *et al.* Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol* **2**, e379 (2004).
- Bettencourt-Dias, M. *et al.* Genome-wide survey of protein kinases required for cell cycle progression. *Nature* **432**, 980–987 (2004).
- Sveicz, A., Novak, B. & Mitchison, J. M. Size control in growing yeast and mammalian cells. *Theor. Biol. Med. Model.* **1**, 12 (2004).
- Castedo, M. *et al.* Cell death by mitotic catastrophe: a molecular definition. *Oncogene* **23**, 2825–2837 (2004).
- Huang, J. Y. & Raff, J. W. The dynamic localisation of the *Drosophila* APC/C: evidence for the existence of multiple complexes that perform distinct functions and are differentially localised. *J. Cell Sci.* **115**, 2847–2856 (2002).
- Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
- Cope, G. A. & Deshaies, R. J. COP9 signalosome: a multifunctional regulator of SCF and other cullin-based ubiquitin ligases. *Cell* **114**, 663–671 (2003).
- Hu, J., McCall, C. M., Ohta, T. & Xiong, Y. Targeted ubiquitination of CDT1 by the DDB1-CUL4A-ROC1 ligase in response to DNA damage. *Nature Cell Biol.* **6**, 1003–1009 (2004).
- Tomoda, K., Yoneda-Kato, N., Fukumoto, A., Yamanaka, S. & Kato, J. Y. Multiple functions of Jab1 are required for early embryonic development and growth potential in mice. *J. Biol. Chem.* **279**, 43013–43018 (2004).
- Doronkin, S., Djagaeva, I. & Beckendorf, S. K. The COP9 signalosome promotes degradation of Cyclin E during early *Drosophila* oogenesis. *Dev. Cell* **4**, 699–710 (2003).
- Neufeld, T. P. & Edgar, B. A. Connections between growth and the cell cycle. *Curr. Opin. Cell Biol.* **10**, 784–790 (1998).
- Hay, N. & Sonenberg, N. Upstream and downstream of mTOR. *Genes Dev.* **18**, 1926–1945 (2004).
- Brumby, A. *et al.* A genetic screen for dominant modifiers of a cyclin E hypomorphic mutation identifies novel regulators of S-phase entry in *Drosophila*. *Genetics* **168**, 227–251 (2004).
- Pearson, G. *et al.* Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. *Endocr. Rev.* **22**, 153–183 (2001).
- Garrington, T. P. & Johnson, G. L. Organization and regulation of mitogen-activated protein kinase signalling pathways. *Curr. Opin. Cell Biol.* **11**, 211–218 (1999).
- Manke, I. A. *et al.* MAPKAP kinase-2 is a cell cycle checkpoint kinase that regulates the G₂/M transition and S phase progression in response to UV irradiation. *Mol. Cell* **17**, 37–48 (2005).
- Sanchez, Y. *et al.* Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25. *Science* **277**, 1497–1501 (1997).
- Furnari, B., Rhind, N. & Russell, P. Cdc25 mitotic inducer targeted by chk1 DNA damage checkpoint kinase. *Science* **277**, 1495–1497 (1997).
- Kramer, A. *et al.* Centrosome-associated Chk1 prevents premature activation of cyclin-B-Cdk1 kinase. *Nature Cell Biol.* **6**, 884–891 (2004).
- Stapleton, M. *et al.* A *Drosophila* full-length cDNA resource. *Genome Biol.* **3**, research0080.1–research0080.8 (2002).
- Yandell, M. *et al.* A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **102**, 1566–1571 (2005).
- Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Berkeley *Drosophila* Genome Project, L. Lum and P. A. Beachy for constructs and reagents; T. Mäkelä, M. Laiho, M. Bonke and P. Ojala for critical review of the manuscript, and R. Medema for discussion. This work was supported by the Centre of Excellence in Translational Genome-Scale Biology of the Academy of Finland, Biocentrum Helsinki, University of Helsinki, Sigrid Jusélius Foundation, Finnish Cultural Foundation, Maud Kuistila Foundation and Finnish Cancer Research Organizations.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.T. (jussi.taipale@helsinki.fi) and M.T. (minna.taipale@helsinki.fi).

CORRIGENDUM

doi:10.1038/nature04500

Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks

Michael D. Shapiro, Melissa E. Marks, Catherine L. Peichel, Benjamin K. Blackman, Kirsten S. Nereng, Bjarni Jónsson, Dolph Schluter & David M. Kingsley

Nature 428, 717–723 (2004)

In this Article, we showed that *Pitx1* maps to the major locus controlling pelvic reduction in sticklebacks and shows altered expression, but not altered amino-acid sequence, in a pelvic-reduced population. Owing to a technical error, Figs 3 and 4 of the Article

incorrectly show results from a *Pitx1* sense probe, rather than from an antisense probe. Similar staining is not observed for sense probes from other genes, raising the possibility that sticklebacks may express some endogenous antisense transcripts from the *Pitx1* region.

A new antisense probe (see Fig. 1, below) shows strong expression in the mouth, lower jaw and developing pelvis of marine larvae. *Pitx1* expression appears normal in the mouth and lower jaw of the Paxton benthic population, but is completely missing in the pelvic region. No expression was detected in neuromasts, thymus, olfactory pits or tail in either population, unlike the pattern seen with the original probe. Despite these differences, our major conclusion, that pelvic reduction results from *cis*-acting regulatory changes in the *Pitx1* locus, remains unchanged. This is because the endogenous sense transcript loses its expression along with the presumptive antisense transcript in the pelvic region of the larval progeny of fish missing the pelvis.

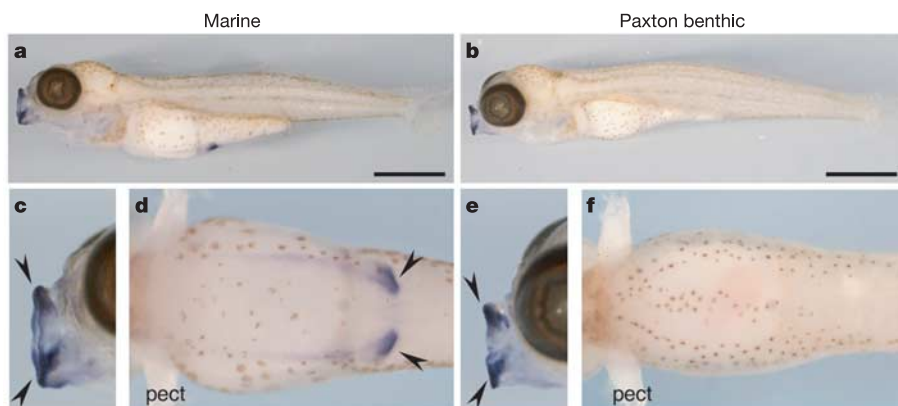


Figure 1 | *Pitx1* is expressed in the prospective pelvic region of marine but not Paxton Lake benthic sticklebacks. **a**, Whole-mount *in situ* hybridization shows *Pitx1* expression in the mouth, jaw and pelvic buds of stage-29 marine larvae. (Details are available from the corresponding author, D.M.K.) **c**, **d**, Enlarged views of the lateral head (**c**) and ventral pelvis (**d**) of stage-29 marine larvae; arrowheads indicate sites of expression. **b**, **e**, **f**, In Paxton benthic larvae, *Pitx1* expression is absent from the prospective pelvic region (**f**), but is present in the mouth and jaw (**e**, arrowheads). Scale bars, 1 mm; pect, pectoral fin.

CORRIGENDUM

doi:10.1038/nature04575

Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy dataAnders Moberg[†], Dmitry M. Sonechkin, Karin Holmgren, Nina M. Datsenko, Wibjörn Karlén & Stein-Erik Lauritzen¹

¹Department of Earth Science, University of Bergen, 5007 Bergen, Norway. [†]Present address: Department of Physical Geography and Quaternary Geology, Stockholm University, 10691 Stockholm, Sweden

Nature 433, 613–617 (2005)

The authorship of this Letter is amended to include Stein-Erik Lauritzen. Details of the Søylegrotta Cave record (series 8), which should have been accredited to S.-E.L., were not supplied in the paper but are available from the corresponding author (A.M., anders.moberg@natgeo.su.se) on request.

In addition, the tree-ring-width data from the Indigirka river region (series G) were inadvertently used without the proper permissions: although the series has been discussed in the literature¹, they are unpublished data that have not been made publicly available; they may, however, be obtained through A.M.

1. Sidorova, O. V., Naurzbaev, M. M. Response of *Larix cajanderi* to climatic changes at the Upper Timberline and in the Indigirka River Valley [in Russian]. *Lesovedenie* 2, 73–75 (2002).

CORRIGENDUM

doi:10.1038/nature04573

Genomic perspectives in microbial oceanography

Edward F. DeLong & David M. Karl

Nature 437, 336–342 (2005)

It has been drawn to our attention (by J. A. Fuhrman) that Fig. 2 contains a citation error. Specifically, the citation associated with “Discovery of planktonic marine archaea” in Fig. 2 that was incorrectly given as ref. 20 should instead be ref. 21 (Fuhrman, J. A., McCallum, K. & Davis, A. A. *Nature* 356, 148–149; 1992).

Fast forward

Proteomics raises new challenges in protein purification. Technologies well adapted to isolate individual proteins get a makeover to tackle large numbers of samples. Laura Bonetta investigates.

Proteomics aims to identify the cellular functions of all proteins encoded by the genome of an organism. Protein structure determination, proteome-wide functional screens, and the identification of protein interactions are just a few of the proteomics applications requiring thousands of purified proteins as a starting point. Researchers now have many reagents and off-the-shelf kits for the isolation of proteins to 90% purity. The challenge facing researchers is to adapt these methods to purify hundreds of different proteins in parallel using robotic systems. Several years ago a small number of academic institutions and biotechnology companies responded to the challenge by developing pipelines for high-throughput protein purification (see "Proteomics at Harvard"). Some of the reagents and instruments required for these methods are making their way to the marketplace.

HT purification

One widely used method for purifying hundreds of proteins in parallel involves their expression in a heterologous system. Once a target protein is identified, the corresponding

complementary DNA is cloned into an expression vector for producing the protein in *Escherichia coli* or another organism. The purification of the expressed recombinant protein from bacteria typically involves cell lysis, incubation of the lysate with an affinity resin, washes and elution. Additional purification steps, such as ion-exchange and size-exclusion chromatography, are required for protein crystallography (see "Protein Purification for Structural Proteomics").

One bottleneck in high-throughput protein production is the expression of enough properly folded protein in bacterial cells. Proteins expressed in *E. coli* can accumulate as inactive, misfolded aggregates called inclusion bodies. Proteins in this form can be purified under denaturing conditions, but need to be refolded into their native conformations. A number of groups have developed procedures for expressing proteins in insect cells or cell-free extracts to get around some of these problems (see "Cell or Cell-Free?").

Each protein is expressed in different amounts and has different properties. Thus, to be able to apply the same purification protocol



PALL CORPORATION

Splitter: Pall's AcroPrep 96 Filter Plate.

across the range of proteins, researchers have engineered proteins with generic tags that will bind to an affinity ligand. Widely used tags include a small peptide of six histidine residues (6xHis), a calmodulin-binding peptide, the streptavidin friendly biotin, the cellulose-binding tag, the maltose-binding protein (NusA), and glutathione S-transferase (GST). Vectors designed for expressing tagged proteins can be purchased from several manufacturers includ-

PROTEOMICS AT HARVARD

Founded in the spring of 1999, the Harvard Institute of Proteomics (HIP) at Harvard Medical School aims to provide tools to determine the function of every protein encoded by the human genome and appropriate model and disease organisms. To this end, HIP scientists are building collections of genes from humans and organisms including *Saccharomyces cerevisiae*, *Vibrio cholerae*, *Yersinia pestis*, *Pseudomonas aeruginosa* and *Bacillus anthracis*, as well as some mouse and viral genes. In addition, a number of projects are focused on specific groups of medically relevant genes.

The Breast Cancer 1000 Project has developed a repository of clones for 1,000 genes that contribute to the onset of breast cancer, a subset of which have already been tested in functional assays to identify cDNAs that induce cancer-like phenotypes.



Another project aims to clone the entire range of human kinases. The expression clones generated at HIP, along with the technology to use them, will be made available to all researchers.

Joshua LaBaer (pictured), co-founder and director of the institute, says his group has developed methods for the high-throughput purification of proteins

expressed in *Escherichia coli* and downstream processes to determine their functions or properties. "In one project, we are purifying all the proteins produced by the organism *Francisella tularensis*, the parasite that causes tularaemia," says LaBaer. The organism's proteome consists of about 1,600 proteins, all of which have been expressed in *E. coli* and purified. The purified proteins are then used in high-throughput functional screens "to find proteins that produce an immune response". A similar study focuses on the genome of the bacterium *V. cholerae*, which can cause cholera in humans. Such studies are not only yielding important drug targets but laying the groundwork for studies targeting the much more complex human proteome.

LaBaer says that two bottlenecks affecting the purification of proteins from

Y. pestis, the plague bacterium, are the "the ability to express large or hydrophobic proteins in bacteria, and avoiding the inclusion-body problem." Partly for this reason the institute is now moving towards the development of protein arrays, with proteins being synthesized on the array rather than spotted on them. In their protocol, plasmid DNA is spotted on the array and genes are then transcribed and translated in a cell-free system. The resulting proteins, hundreds of them per chip, are immobilized *in situ* and can be used to test for protein-protein interactions and other functional assays. Although the technology is not yet ready for primetime, LaBaer says they are having success with it. The development of these types of arrays may alleviate the need to produce and purify proteins from *E. coli*, as scientists will be able to conduct functional studies directly on the arrays.

L.B.

ing Invitrogen (Carlsbad, California), Novagen (Madison, Wisconsin), Roche Applied Science (Indianapolis, Indiana) and others.

Some companies are now developing high-throughput methods for purifying proteins that are not tagged. "Typically most scientists work with tagged proteins through the discovery process. However, it is possible to lose some information because of tag interference in the biological assays. Removing the tag means removing some of the question marks," says Lisa Bradbury, director of R&D for proteomics and cell therapy at Pall Corporation of East Hills, New York.

Purification tool bag

Several companies sell reagents for every step of the purification process — from cell lysis, to affinity purification resins, to desalting and concentrating reagents — adapted to high-throughput protocols that use standard 96-well plates. "We were recognized for our protein affinity-purification line. Now we are configuring it into a format that is automation-friendly," says Craig Smith, vice-president of R&D at Pierce Biotechnology of Rockford, Illinois (part of Fisher Scientific). Among the company's offerings are the SwellGel Discs, pellets of dehydrated support that can be distributed into filter plates. When the protein is



High-capacity purification: Teledyne Isco's BioOptix 10.

added, the disc rehydrates to an affinity gel that binds tagged proteins. "We have taken out variability by having a dried pellet," says Smith. More recently, the company's Zeba Micro Desalt Spin Columns were configured to 96 well-plate format for desalting many small samples in parallel.

Pall Corporation also sells reagents for each step of the purification process. The AcroPrep 96 Filter Plate can be used for any type of chromatography, including immobilized metal affinity chromatography, in a multiwell platform. The format facilitates the optimization of various purification parameters, including the choice of metal ion and resin, the sample-to-resin ratio, and the elution conditions. Other reagents adapted to a high-throughput

environment include the Mustang 96-well ion-exchange plates and Nanosep Centrifugal Devices.

QIAGEN, based in Hilden, Germany, offers the Superflow 96 BioRobot, a medium-scale purification kit for 6xHis-tagged proteins. "You load your pellets onto a robot, press a button, and you will have product in two hours," says Frank Schäfer, associate director for R&D protein expression/proteomics. According to the company, purification can be done in denaturing or native conditions and yields 4 mg of protein for each of 96 wells. "It is one system and one process for every sample," says

Schäfer. QIAGEN sells a similar system that runs up to 24 samples in parallel to purify up to 30 mg of protein for large-scale applications.

Novagen's host of reagents for high-throughput protein purification starts with the expression step. The Overnight Express Autoinduction Systems allow fully automated, parallel protein expression from *E. coli* cultures, and achieve high cell densities without monitoring growth or induction by the addition of isopropyl-beta-D-thiogalactopyranoside. "The medium contains a blend of carbon sources optimized for tightly regulated, uninduced growth and automatic induction of protein expression," says Anthony Grabski, R&D group leader for protein purification. The RoboPop Purification Kits contain reagents for the extraction and purifica-

PROTEIN PURIFICATION FOR STRUCTURAL PROTEOMICS

Structural genomics, or structural proteomics, aims to provide three-dimensional information for all proteins. This information can be used to ascribe function to a protein and to reveal or invalidate drug targets. There are structural proteomics projects in both the public and private sectors around the world, including Germany, Japan and the United States.

In 2000, the US National Institute of General Medical Science in Bethesda, Maryland, launched an initiative to determine the structures of about 10,000 proteins representing different structural ('fold') families over the next decade. The studies will result in a public resource linking amino-acid sequence, structural and functional information, eventually allowing scientists to make the three-dimensional atomic structures of most proteins easily obtainable from their corresponding sequences.

One of several participating centres is the Genomics Institute of the Novartis Research Foundation in San Diego, a member of the Joint Center for Structural Genomics (JCSG). Scott Lesley (pictured below, left), who heads the institute's proteomics unit, says his group has processed 50,000 protein samples from more than 20,000 unique expression clones. The pipeline starts with a small screen protocol

that uses samples in a 96-well format to evaluate how the targets behave at different steps in the procedure. "Those that are not behaving well after expression, purification or crystallization are set aside and then run through one of several salvage pathways," says Lesley. Such pathways include using different expression systems (vectors and cells), making point mutations in the cDNA to increase the likelihood it will be expressed, and trying different tags. "All the bacterial expression and purification is done in parallel and the majority is automated by custom instrumentation," says Lesley.

The targets that perform well are applied to a large-scale purification protocol to yield

10–20 mg of protein. The JCSG has already deposited 239 structures in the protein database — mainly from the bacterium *Thermotoga maritima*, the mouse, and the yeast *Saccharomyces cerevisiae* — representing at least 19 new fold families. "Most of the tools are in place, and we can increase both output and success rate by implementing salvage approaches in high throughput," says Lesley.

Similar efforts are underway in other countries. The Protein Structure Factory in Germany, an initiative of the German Human Genome Project and structural biologists from the Berlin area, targets human proteins for structure determination. "Our current focus is to determine the structures of particular protein-protein or protein-ligand complexes," says Christoph Scheich of the Max Planck Institute for Molecular Genetics in Berlin.



TELEDYNE ISCO

G.N.F.

tion of His or GST fusion proteins by magnetic or filtration-based affinity purification protocols, all in a 96-well format. The kits have been validated on both PerkinElmer and Tecan robotic liquid handlers.

Other providers of reagents for high-throughput protein purification include GE Healthcare of Little Chalfont, UK, with its Tricorn High Performance Columns designed for high-resolution purification of proteins, peptides and other biomolecules. Sartorius of Göttingen, Germany, introduced the Vivapure 8-to-96 well cobalt chelate kit for the simultaneous purification of multiple His-fusion proteins by affinity chromatography. And St Louis-based Sigma-Aldrich's HIS-Select iLAP Plates are plates coated with cell-lysis reagents and a HIS-Select nickel chelate matrix allowing for cell lysis, protein capture, and assay of a His-fusion protein in a single well.

As well as purification kits, several companies offer products to help screen the outcome of each step in the purification pipeline. Novagen's RoboPop Solubility Screening Kit contains a filtration plate that retains insoluble inclusion bodies while allowing soluble proteins to be collected for rapid quantitation and analysis. "You can screen expression conditions for soluble protein before you proceed to purification," says Grabski. In January, the company released the iFold Protein Refolding System 1 to screen different refolding conditions in parallel. The system includes inclusion-body purification reagents and 92 different refolding buffers. "Structural proteomics

groups have harvested most of the low-hanging fruits, so now they will have to focus on difficult proteins that are insolubly expressed," says Grabski. Another product for monitoring the purification procedure is the Protein 200-HT2 assay of Agilent Technologies (Palo Alto, California), which allows the identification, sizing and quantification of proteins from 14 kD to 200 kD in size. "It allows researchers to replace SDS-PAGE procedures in the lab," says product manager Carsten Buhlmann. Finally, some reagents and instruments have been designed to help researchers hone in, for example, on medically relevant targets, which can then be purified and studied (see "Target Identification").

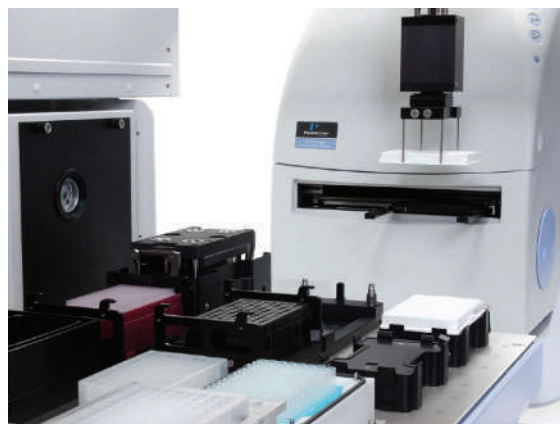
Getting automated

The available reagents and kits can be used to purify several hundred proteins in a 96-well format at considerable speed. However, the tedious and error-prone nature of manually performed high-throughput operations calls for automation of the process. "When a robot is doing the pipetting there are no mistakes, so you can have greater consistency and reproducibility," says David Daniels, applications marketing manager for Beckman Coulter in Fullerton, California. Liquid handlers perform all the steps of protein purification from loading cell cultures to obtaining a protein in solu-

tion, in under four hours.

Two leading liquid handlers are Beckman Coulter's Biomek 3000 and Biomek FX instruments, with plate-deck configurations that can handle 10 and 18 plate positions, respectively. The Biomek 3000 will process two 96-well plates in 2–3 hours; the FX twice as many.

Promega in Madison, Wisconsin, automated high-throughput protein purification using its MagneHis Protein Purification System and the Biomek FX instruments. The MagneHis reagent contains a cell-lysis solution that allows resuspension and lysis of bacterial cell pellets without sonication and centrifugation. Magnetic pre-charged nickel particles are then used to isolate His-tagged proteins from the cell lysate. The MagneHis particles bound to the target proteins are captured on a MagnaBot 96



Mix and match: PerkinElmer's JANUS workstation.

PERKINELMER

CELL OR CELL-FREE?

Only a fraction of proteins can be overproduced in *E. coli* in sufficient yield and without the formation of inclusion-body aggregates or the proteolytic degradation of expressed proteins. Alternative expression systems include cell cultures from eukaryotic organisms, such as insect cells, and cell-free, *in vitro* protein expression.

The latter is the focus of John Markley's group at the Center for Eukaryotic Structural Genomics at the University of Wisconsin, Madison. Collaborating with Ehime University and CellFree Sciences, both in Japan, Markley and colleagues developed a pipeline using wheat germ cell-free protein translation as a way to produce proteins for nuclear magnetic resonance (NMR) structure determination.

Cell-free systems simplify the purification efforts, as "only the protein of interest is expressed



and labelled, thus the background is cleaner", says Markley (pictured). For the NMR studies, his group has been getting twice as many folded proteins using the cell-free system than with expression in *E. coli*. The cell-free method also requires smaller volumes, avoiding lengthy concentration steps, and it lends itself to the labelling of proteins,

which is a requirement for NMR. "But there is a steep learning curve for learning how to do a cell-free system," says Dmitriy Vinarov, who is responsible for high-throughput production at the centre. One downside of cell-free systems is the expense of the reagents, especially when success rates are not high. "About 79% of human proteins will be expressed,

about one half of those will produce protein in sufficient quantities for structural studies, and half of those will remain stably folded for NMR studies," says Markley.

Cell-free and cell-based systems are not mutually exclusive. "Each has unique advantages and capabilities so we will continue to do both," explains Markley. By using a new cloning system from Promega the researchers have been able to transfer target cDNAs from cell-free to cell-based expression systems to achieve greater flexibility. "We still have a lot to learn. It is working quite well for us but we still have improvements to make," says Vinarov. "The technology is not quite primetime."

Cell-free systems are available from companies such as Roche Applied Science, QIAGEN and Invitrogen, as well as from CellFree Sciences.

R. DAVIES

Magnetic Separation device — a plate that can be hooked up to the Biomek instruments.

In January, PerkinElmer (Boston, Massachusetts) launched its modular and scalable JANUS Automated Workstation. “Many units can be configured at the time of sale,” says Nance Hall, business unit leader for automation and liquid handling. “With JANUS, you can scale and change the system not only when you buy, but also in the future.” Users can also program the selection of numerous dispense heads and formats. “It is as though you go into a kitchen and you are asked if you need a cup or a teaspoon,” explains Hall. “In the kitchen you will use both.” Likewise, the JANUS allows users to choose different tools automatically depending on the volume range and microplate densities.

Another popular choice among liquid handlers is the Freedom EVO automated liquid handler from Tecan in Durham, North Carolina. Others include the Sciclone ALH Workstation from Caliper Sciences (Hopkinton, Massachusetts), the BioCube System from Proteodyne (Windsor, Connecticut) and the 925 PC Workstation and GX-281 Liquid Handler from Gilson (Middleton, Wisconsin). “You can use over 300 standard racks and many more custom ones,” says Gilson’s spokesperson Greg Robinson.

Going large

For crystallization and other functional studies, a researcher needs more protein than is possible in 96-well plate format. For such purposes, a number of parallel-purification sys-



Just swell: Pierce Biotechnology's SwellGell discs.

tems that can accommodate larger volumes of cells are starting to come onto the market.

GE Healthcare's AKTExpress is a fully automated chromatography system for purification of His- and GST-tagged proteins, yielding up to 50 mg of target protein. Affinity chromatography is the first step of all protocols, but as a second step it is possible to choose between desalting or gel filtration, and an ion-exchange step can be added too. Automatic tag removal is possible in all multi-step protocols.

Teledyne Isco (Lincoln, Nebraska) launched the BioOptix 10 last year, a ten-channel parallel-purification system for high-capacity protein purification with subsequent fraction collection. This means it can purify protein

from ten different samples by ion exchange, affinity or size-exclusion chromatography. The instrument includes a high-capacity fraction collector (20 to 60 fractions per sample). Ten independently controlled pumps can be programmed with different gradient conditions for rapid identification of columns and conditions that give optimal results. “Instead of doing it sequentially, a process that can take 4–5 days, you can load different columns and run the instrument over lunch,” says John Urh, product manager for chromatography.

Another player in this arena, QIAGEN's BioRobot Protein LS System, has the capacity for the parallel purification of up to 24 large-scale cultures in less than three hours, and PerkinElmer's JANUS system can be configured for large-scale purification protocols.

New reagents and instruments have allowed researchers to set up pipelines for high-throughput protein production of a scale and capacity that match the needs of their individual labs. As a result of these advances, hundreds of proteins from different organisms have been purified and their structures and functions determined. But keeping abreast of this rapidly changing field will require ongoing innovation. “Several years ago everyone was talking about genomics, now the focus is on proteomics. More and more we are seeing research moving toward specific proteins,” says PerkinElmer's Hall. “The challenge is trying to keep up with the technologies.”

Laura Bonetta is a freelance writer based in the Washington DC area.

TARGET IDENTIFICATION

High-throughput protein-purification pipelines are becoming part of many proteomics efforts. But how to choose the targets to put through the pipeline? For those who want to focus on targets relevant to a disease-related process, it is important to catalogue where, when and to what extent a protein is expressed. Currently the main method for determining the protein complement of a given cell or tissue uses two-dimensional (2D) gel electrophoresis or liquid chromatography to resolve the proteins in the sample, and mass spectrometry to then identify the individual proteins. A problem with this approach is that it is difficult to identify rare proteins, because cell extracts are dominated by a very few abundant proteins.

Beckman Coulter has developed a number of innovations for protein fractionation to address

this problem. “We first selectively remove the most abundant proteins from a sample,” says Jerry Feitelson, marketing manager for the company. The ProteomeLab IgY-12 partitioning kits selectively remove the 12 most abundant proteins — albumins and immunoglobulins — from human/primate serum or plasma. These proteins make up 95% of the total protein expressed by cells. So removing them increases the chances of identifying proteins that are not highly expressed and are probably more interesting to researchers. The technology uses antibodies bound to inert beads, which are then packed in liquid-chromatography columns or, for smaller samples, spin columns. “The advantages of our reagents are the greater capacity, increased specificity and the ability to bind across species,” says Feitelson. The enriched material can then be collected and further fractionated

with the Beckman Coulter ProteomeLab PF 2D protein fractionation system. The instrument divides proteins in two dimensions: first, in a 2D gel, samples are separated by isoelectric focusing and size, and then automatically injected into a liquid-chromatography column. The process is fully automated. “One run takes ten hours, so you can do two in one day,” says Feitelson. Detailed protein maps can be constructed for easy comparison between two samples using the Proteome Lab software suite. Interesting proteins can then be further analysed by cutting out the corresponding band, for example, and putting it into a mass spectrometer.

Pall Corporation has also introduced the Enchant Life Sciences kits for albumin depletion and immunoglobulin G purification.

L.B.



H. DEWALD/BECKMAN COULTER

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Affinity tags purification			
Active Motif	Ni-TED Protein Purification System, spin columns	Carlsbad, California	www.activemotif.com
Alpha Diagnostic International	Antibodies to a number of common fusion-tags	San Antonio, Texas	www.4adi.com
BD Biosciences	Affinity resins for the purification of tagged proteins	San Diego, California	www.bdbiosciences.com
Bio-Rad	Profinity IMAC resin, protein extraction kits	Hercules, California	www.bio-rad.com
Biotech Support Group	NuGEL affinity purification, HPLC columns, serum protein analysis tools	North Brunswick, New Jersey	www.biotechsupportgroup.com
Clontech (Takara Bio)	GST and His fusion protein purification kits, affinity resins	Mountain View, California	www.clontech.com
● Covalys Biosciences	SNAP-capture kits for SNAP-tag fusion proteins	Witterswil, Switzerland	www.covalys.com
IBA	Strep-Well HT Purification plates for automated purification of streptavidin-tagged proteins	Göttingen, Germany	www.iba-go.com
● Invitrogen	Ni-NTA and ProBond purification systems for the purification of His ₆ -tagged proteins	Carlsbad, California	www.invitrogen.com
Macherey-Nagel	Products for chromatography and sample preparation	Düren, Germany	www.macherey-nagel.com
MoBiTec	Columns and kits for protein purification	Göttingen, Germany	www.mobitec.de
NeoClone	Softag epitope purification system	Madison, Wisconsin	www.neoclone.com
● Novagen (EMD)	Protein purification affinity resins, Robopop purification kits, magnetic beads	Madison, Wisconsin	www.novagen.com
Pall Life Sciences	Automation-friendly multi-well filter plates for cleanup, purification, and detection of proteins, reagents for sample cleanup and concentration	East Hills, New York	www.pall.com
Pierce Biotechnology	Kits for GST- and His-fusion proteins, purification columns, affinity resins, SwellGel discs	Rockford, Illinois	www.piercenet.com
Phenomenex	Chromatography columns	Torrance, California	www.phenomenex.com
PhyNexus	PhyTip IMAC and glutathione products	San Jose, California	www.phynexus.com
Prochem	Proteus spin columns for the purification of His ₆ -tagged recombinant proteins	Littleton, Massachusetts	www.protein-chem.com
Profos	ColTrap Proteininnative system for purification of His-tagged proteins expressed in <i>E. coli</i> under native conditions	Regensburg, Germany	www.profos.de
● R&D Systems	His-tagged protein purification resin	Minneapolis, Minnesota	www.rndsystems.com
Rockland Immunochemicals	Antibodies with GST and GFP epitope tags, fluorescein and peroxidase antibodies, alkaline phosphatase and biotin conjugates	Gilbertsville, Pennsylvania	www.rockland-inc.com
Sartorius	Vivapure 8-to-96-well cobalt-chelate kit, ion exchange columns	Göttingen, Germany	www.sartorius.com
● Sigma-Aldrich	Affinity purification plates and gels, reagents for cell lysis and protein precipitation, His-Select HC nickel magnetic beads	St Louis, Missouri	www.sigmaaldrich.com
Stratagene	InterPlay TAP kits	La Jolla, California	www.stratagene.com
Upstate	Spin column filters	Charlottesville, Virginia	www.upstate.com
● QIAGEN	Ni-NTA Superflow BioRobot, MagneHis Protein Purification System	Hilden, Germany	www.qiagen.com
Tosoh Biosep	Columns and media for chromatography	Montgomeryville, Pennsylvania	www.tosohbiosep.com
ZirChrom Separations	Chromatography columns	Anoka, Minnesota	www.zirchrom.com
Zymo Research	His-Spin Protein Miniprep for His-tagged protein purification	Orange, California	www.zymoresearch.com
Magnetic beads and magnetic separators			
● Ambion	Magnetic Stand-96 for paramagnetic bead precipitation	Austin, Texas	www.ambion.com
Bioclone	Recombinant proteins, magnetic beads, magnetic separators	Marrickville, Australia	www.bioclone.com.au
Bio-Nobile	QuickPick protein fractionation and purification, PickPen magnetic tool	Turku, Finland	www.bio-nobile.com
Bioscience Bead Division	Magnetic beads	West Warwick, Rhode Island	www.bioscience-beads.com
Bruker Daltonics	CLINPROT magnetic beads for desalting and quick sample preparation	Billerica, Massachusetts	www.bdal.com
Invitrogen	Dynabeads TALON for the isolation of His-tagged proteins	Carlsbad, California	www.invitrogen.com
Europa Bioproducts	Recombinant proteins, expression-ready clones, magnetic particles, BcMag magnetic separator	San Diego, California	www.bioclon.com
Magnetic Biosolutions	Regenerable magnetic streptavidin beads	Stockholm, Sweden	www.magbio.com
Miltenyi Biotec	Paramagnetic MACS MicroBeads, ProCatch Resins, magnetic separators	Auburn, California	www.miltenyibiotec.com
New England Biolabs	Streptavidin magnetic beads, magnetic separators	Ipswich, Massachusetts	www.neb.com
Promega	MagneGST Protein Purification System	Madison, Wisconsin	www.promega.com
Polysciences	BioMag superparamagnetic particles	Warrington, California	www.polysciences.com
StemCell Technologies	EasySep magnet	Vancouver, Canada	www.stemcell.com
Thermo Electron Corporation	KingFisher automated system for magnetic particle-based protein purification	Waltham, Massachusetts	www.thermo.com
Chromatography equipment and automated systems			
Artel	Multichannel verification system portable technique to verify the performance of liquid-handling equipment	Westbrook, Maine	www.artel-usa.com
Aurora Biomed	VERSA robotic workstation	Vancouver, Canada	www.aurorabiomed.com
Agilent Technologies	ZORBAX HPLC columns for protein and peptide separations, combined capillary electrophoresis and mass spectrometry system	Palo Alto, California	www.agilent.com
● Applied Biosystems	Protein sequencers, mass spectrometers, chromatography systems	Foster City, California	www.appliedbiosystems.com
Beckman Coulter	Biomek liquid handlers for HT protein purification, ProteomeLab PF2D for protein fractionation	Fullerton, California	www.beckmancoulter.com
Biotage	FLASH chromatography systems for multi-sample purification	Uppsala, Sweden	www.biotage.com
Bio Tek Instruments	Complete line of 96- and 384-well microplate automated pipetting systems, dispensers and washers	Winooski, Vermont	www.biotek.com
Bruker Daltonics	Mass spectrometry instruments and accessories for pharmaceutical, biochemical and chemical research	Billerica, Massachusetts	www.bdal.com
Caliper Life Sciences	Advanced liquid handling and LabChip technologies	Hopkinton, Massachusetts	www.caliperLS.com
Cecil Instruments	Products for ion chromatography, HPLC and spectrophotometers	Cambridge, UK	www.cecilinstruments.com

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Eksigent Technologies	ExpressLC-800 HPLC system, NanoLC-1D and NanoLC-2 proteomics systems for nanoscale chromatography	Livermore, California	www.eksigent.com
● GE Healthcare	AKTExpress automated liquid chromatography platform	Little Chalfont, UK	www.gehealthcare.com
Gilson	Solid phase extraction systems, liquid chromatography	Middleton, Wisconsin	www.gilson.com
Gyros	Gyrolab Workstation for automated sample preparation for mass spectrometry	Uppsala, Sweden	www.gyros.com
Hamilton Company	Automated solid phase extraction systems	Reno, Nevada	www.hamiltoncompany.com
Hitachi High-Technologies	Process Automation System can automate various laboratory processes ranging from centrifuging to sample partitioning and analysis	Tokyo, Japan	www.hitachi-hitec.com/oversea
PerkinElmer Life Sciences	Range of liquid handlers and automated workstations	Boston, Massachusetts	las.perkinelmer.com
Proteodyne	BioCube automated protein expression and purification	Windsor, Connecticut	www.proteodyne.com
Seplatec	Automated HPLC and solid phase extraction systems for sample separation and isolation	Berlin, Germany	www.seplatec.com
Tecan	Freedom EVO workstation, solid-phase extraction systems	Männedorf, Switzerland	www.tecan.com
TTP LabTech	Mosquito liquid-handling robot for setting up protein crystallography screens	Royston, UK	www.ttplabtech.com
Varian	Chromatography and mass spectrometry analytical instruments and consumables	Walnut Creek, California	www.varianinc.com
Waters	Products and tools for liquid chromatography and mass spectrometry	Milford, Massachusetts	www.waters.com
Protein concentration, desalting and extraction kits			
Ambion	PARIS kit for the isolation of both RNA and native protein from the same sample	Austin, Texas	www.ambion.com
BioChain	Protein extraction kits	Hayward, California	www.biocchain.com
BioVision	Reagents for preparing highly enriched cellular fractions from mammalian cells	Mountain View, California	www.biovisionlabs.com
Calbiochem (EMD)	Protein extraction and concentration kits	San Diego, California	www.emdbiosciences.com
Chemicon	Protein extraction and concentration kits	Temecula, California	www.chemicon.com
Eppendorf	Perfect Pure tips for sample preparation for mass spectrometry, centrifuges	Hamburg, Germany	www.eppendorf.com
EPICENTRE Biotechnologies	ReadyPreps preparation kit for lysing bacterial cells and reducing viscosity for protein extraction	Madison, Wisconsin	www.epibio.com
Gerard Biotech	Protein concentration kits for denatured and native proteins	Oxford, Ohio	www.gerardbiotech.com
Norgen Biotek	ProteoSpin kit for concentration, buffer exchange and desalting of protein samples	St Catharines, Canada	www.norgenbiotek.com
Novagen (EMD)	iFOLD protein refolding system	Madison, Wisconsin	www.novagen.com
Profos	ColTrap Protein Unfold and Protein Express, endotoxin removal, ColTrap: bacteria harvesting	Resensburg, Germany	www.profos.de
Sigma-Aldrich	ProteoPrep protein extraction kits	St Louis, Missouri	www.sigmaaldrich.com
● Takara Mirus Bio	Kits for optimizing the refolding conditions of inclusion body proteins	Madison, Wisconsin	www.takaramirusbio.com
Services			
Abgent	Bacterial and baculovirus protein expression systems	San Diego, California	www.abgent.com
Arvys Proteins	Complete protein expression and purification service	Stamford, Connecticut	www.arvysproteins.com
BioGenes	Preparation and purification of antigens and development and production of antibodies	Berlin, Germany	www.biogenes.de
Hypromatrix	Protein-protein interaction assays, protein purification	Worcester, Massachusetts	www.hypromatrix.com
Invitrogen	Customized protein expression, purification and characterization	Carlsbad, California	www.invitrogen.com
Nature Technology	Contract manufacturing of vectors (DNA, viruses, vaccines, plasmids) and recombinant proteins	Lincoln, Nebraska	www.natx.com
Paragon Bioservices	Recombinant protein expression and purification services	Baltimore, Maryland	www.paragonbioservices.com
Protagen	Protein production and analysis services	Dortmund, Germany	www.protagen.de
Protein Sciences	Cloning, expression and purification of full-length, biologically active proteins	Meriden, Connecticut	www.proteinsciences.com
ProteinX Lab	Contract protein purification services	San Diego, California	www.proteinx.com
● Roche Applied Science	Protein production by cell-free and cell-based expression systems	Indianapolis, Indiana	www.roche-applied-science.com
Trenzyme Biotechnology	Protein expression and purification services	Konstanz, Germany	www.trenzyme.com
Other proteomics tools			
Alpha Innotech	Chemiluminescence and gel imagers	San Leandro, California	www.alphainnotech.com
AnaSpec	Protein detection kits	San Jose, California	www.anaspec.com
Applied Biosystems	BIOiTRAQsystems for protein biomarker identification and quantitation	Foster City, California	www.appliedbiosystems.com
Bio-Rad	Preparative 2D electrophoresis, isoelectric focusing systems, proteomics workstations	Hercules, California	www.bio-rad.com
Ciphergen	Protein chip systems for biomarker discovery	Fremont, California	www.ciphergen.com
Eksigent Technologies	Microfluidics-based systems for proteomics analysis	Dublin, California	www.eksigent.com
Eurogentec	Production of custom protein arrays	Seraing, Belgium	www.eurogentec.be
Genomic Solutions	Instrumentation, software, and consumables to isolate, identify and characterize proteins	Ann Arbor, Michigan	www.genomicsolutions.com
JEOL	Magnetic sector mass spectrometers	Tokyo, Japan	www.jeol.com
NextGen Sciences	a2DEoptimizer 2D gel electrophoresis system	Huntingdon, UK	www.nextgensciences.com
Protein One	Active Protein Array platforms designed for multiplex detection and interaction assessment with DNA, RNA, and/or other ligands	Bethesda, Maryland	www.proteinone.com
Proteome Systems	ElectrophoretIQ for first-dimension isoelectric focusing and second dimension SDS-PAGE and blotting in a single bench top unit	Woburn, Massachusetts	www.proteomesystems.com
Proxeon	Mass spectrometry tools and products	Odense, Denmark	www.proxeon.com
Thermo Electron Corporation	Finnigan ProteomeX Integrated Proteomics Workstation for protein and peptide analysis and identification	San Jose, California	www.thermo.com/chromatography
Whatman	FAST PAK protein array kit line	Brentford, UK	www.whatman.com

● Advertiser

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Having it all

Grad students and postdocs know about career opportunities in sectors beyond academia. But they are uncertain about how to pursue them, and are wary about leaving academia and getting stuck in a sector that, they think, offers no return to academic science. A group of young scientists attending a career seminar at a meeting sponsored by Nature Biotechnology and the University of Miami voiced those common concerns earlier this month.

The students were encouraged by one panellist's career path. Jeremy Paul, director of the Frontiers of Science Program at the New York Academy of Sciences (NYAS), has worked in academic, industrial and non-profit sectors, both on and off the bench. After a postdoc at the University of Chicago, Illinois, where he studied the structure and function of recombinant human insulin receptors, he helped, with a group of academic scientists, to found Progenics Pharmaceuticals in Tarrytown, New York. From there, Paul moved on to Cadus, another biotech company in upstate New York. There, he learned how to manage interactions both with big pharma and non-profits, as the company had alliances with each. Next, he took his growing

project-management experience to OSI Pharmaceuticals, based in Farmingdale, New York, and expanded his expertise in intellectual property, before moving on to Aton Pharma, another upstate New York biotech, where intellectual property became his sole focus. When he joined NYAS last year, Paul applied his project-management skills in a different sector. He also drew on his experience in communicating across different sectors to help the NYAS build up its educational offerings.

Paul's path shows that one can start off in academia, work in industry, get involved in non-profit work, and remain engaged in academic science — as long as you stay in contact with people from earlier in your career, build up skill sets and learn how to apply existing ones in new environments.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building,
 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)
Natureevents: Silke Opstrup (4994)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
**Germany/Austria/
 The Netherlands:**
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

**European Satellite Office
 Germany/Austria/
 The Netherlands:**
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Pacific sunrise

The Pacific Northwest of North America doesn't just mean Microsoft, Intel and some big trees. Already noted for the quality of its biological research, the biotechnology base in cities such as Vancouver is set to grow too, as **Virginia Gewin** finds out.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Stretching from Vancouver in British Columbia to Oregon, the Pacific Northwest may not be the richest or the largest technology cluster in the world, but it has a flavour all its own.

Home of the Bill and Melinda Gates Foundation, with its commitment to global health, the city of Seattle brings an idealistic spirit to science and technology. Add some world-class medical and scientific institutions, and the region is fast developing into an influential centre for the life sciences. The public sector still offers the greatest number of jobs, but the private sector is growing.

The Pacific Northwest of the United States and Canada — Washington, Oregon and British Columbia — has given birth to industry leaders. From Microsoft to Amazon, technology is key. It comes as no surprise that the medical charity established by Microsoft co-founder Bill Gates is one of the driving forces of the region's current biomedical boom.

"The importance of the Gates Foundation and its emphasis on global health cannot be overstated," says Nancy Haigwood, viral vaccines programme director



Bill Gates: foundation funding is driving the boom.

at the Seattle Biomedical Research Institute (SBRI). Non-profit organizations strong in immunology and vaccine research, such as SBRI, have received significant chunks of the Gates' Grand Challenge grants. SBRI alone has been awarded Gates grants of \$47.5 million since 2000. That influx of funds is still producing job opportunities, particularly in the field of infectious diseases. In 2000, there were only five malaria researchers in the SBRI's Malaria Antigen Discovery programme. Now, there are more than 60 — with more due to be recruited.

Headquartered in Seattle, the non-profit organization Path focuses on promoting sustainable international health technologies. As a result of Gates funding it has 60 open positions worldwide in areas as diverse as microbiology to vaccine development. The people at Path are also excited by the recently launched Gates-funded Department of Global Health at the University of Washington in Seattle. The university is a powerful force in the northwest and is the largest life-sciences employer in the region, followed closely by the Fred Hutchinson Cancer Research Center.

At the Seattle-based Institute of Systems Biology (ISB), the non-profit research institute founded by sequencing innovator Leroy Hood, a \$13-million grant from Gates and California-based pharmaceutical giant Amgen will help fund research on the prediction and prevention of disease.

The Allen Institute of Brain Science in Seattle, set up by Microsoft's other founder, Paul Allen, is set to finish the \$100-million Allen Brain Atlas this year. The institute will be hiring systems neuroscientists and data managers for the next phase of work — mining the atlas data on brain circuitry and gene expression and an in-depth study of the cortex.

The region's homegrown biotechnology industry has had a less successful time recently, battered by lay-offs and acquisitions such as that of Corixa by GlaxoSmithKline. The area has been noted for

NORTHWEST NANOFUTURES

The Pacific Northwest also aims to make a name for itself in nanotechnology. Set up in 2001, the Pacific Northwest National Lab (PNNL)-University of Washington joint institute for nanoscience has the aim of furthering research and education in the field. The university's Center for Nanotechnology boasts the first PhD programme in nanotechnology in the United States, funded by a National Science Foundation-Integrative Graduate Education and Research Traineeship (NSF-IGERT).

State leaders are building on Oregon's reputation as the 'silicon forest' of semiconductor-based industry to develop nanotechnology. The recently established Oregon Nanoscience and Microtechnologies Institute (ONAMI) is a collaboration between Oregon universities, PNNL and companies including Hewlett Packard at Corvallis and Beaverton-based electronics firm Tectronix. ONAMI recently received \$28 million in state funding to raise additional funds for these fields. The University of Oregon in Eugene is leading the training effort with its new Materials Science Institute, complete with an NSF-IGERT for MS and PhD training at the interface of chemistry and physics.

V.G.

HOTSPOTS

Seattle

Health-related research is expanding in Seattle, with many institutions in the midst of construction to house their growing staff. Starting in 2008, the Washington Life Sciences Discovery Fund will have \$35 million annually to disburse for ten years, potentially creating many new jobs.

California-based pharmaceutical company Amgen plans to take a healthy helping of local talent at its facility in Seattle (pictured below). Richard Farley, Washington director of human resources for Amgen, says it plans to recruit 150 employees over the next year to grow their drug-discovery pipeline — that is a 15% increase in the workforce, primarily in research positions.

Portland

Compared with its neighbours on either side, Oregon's foray into the life sciences could be viewed as baby steps.

But the recent \$500-million public-private investment in Portland's Oregon Health and Sciences University (OHSU) — the Oregon Opportunity — was a tremendous leap towards the state's goal of obtaining its share of the bioscience industry and converting research into clinical reality. The OHSU is halfway through recruiting 70 principal investigators and staff to fill the new 25,500 square metre biomedical research building that will serve as the cornerstone of the project.

Vancouver

Vancouver may be only the eighth largest life-sciences cluster in North America, but its companies offer the highest rate of return to investors. George Hunter, president of Vancouver-based Leading Edge BC, which

promotes technology in the region, credits British Columbia's progressive government policies for their impressive stats. In addition to a tax-credit programme that makes Canada a low-cost place to conduct research, business skills get specialized attention. Leading Edge has developed a one-year mentorship programme called BC Excels to address the skills needed by the biotech community.

Vancouver companies have also shown that they can attract talent in their own right. OncoGenex, which is developing drugs for hard-to-treat cancers, has opened an office in Seattle, having hired some of the clinical team from Corixa, a Seattle-based biotech company that recently merged with pharmaceutical giant GlaxoSmithKline. V.G.



bioinformatics, through companies such as Merck subsidiary Rosetta. But that wave may have crested, according to Randall Schatzman, chief executive of antibody developers Alder Biopharmaceuticals of Bothell, Washington, who suggests that Seattle's 20-year history of protein therapeutics and immunology are what will carry it into the future.

That future includes long-time Seattle protein therapeutics company ZymoGenetics, and big pharma in the form of Amgen and Wyeth. Based in Madison, New Jersey, Wyeth Pharmaceuticals recently signed a deal, potentially worth \$800 million, with Seattle start-up Trubion Pharmaceuticals. They plan to use Trubion's protein-assembly technology to develop a new class of drugs to penetrate tumours and diseased tissue. Many in the region see this as the beginning of a revival for the biotech industry.

Seattle is also taking a new approach to company creation. The ISB serves as the institutional anchor for the Accelerator — a private investment and biotech development company that partners top investors with dedicated management and scientific leaders. The experiment has so far created five companies in its three-year existence.

On the Canadian side of the border, Vancouver also aims to turn university research into industry gold, although like its American neighbours it has suffered from an early dearth of capital investment. "Vancouver is becoming a favoured target for company acquisitions by global multinationals — a place we're happy to be," remarks George Hunter, president of Leading Edge BC, a non-profit organization dedicated to promoting British Columbia as a technology hub. He is excited by the trend for biotech pioneers to become serial entrepreneurs.

Vancouver's prowess in biotech is considerable. Of only 19 profitable biotech companies worldwide, three are in Vancouver. More than 70% of the industry is biopharma, and niche areas, such as oncology, HIV and AIDS research, and bioinformatics, have seen much growth recently, agrees Karimah Es Sabar, executive director of BC Biotech, a regional industry organization. Scott Cormack, chief executive of drug developer OncoGenex, adds that Vancouver is reaching a critical mass necessary for a thriving biotech industry.

One driver of Vancouver's biotech is a heavy focus on genome sciences, particularly research into disease biomarkers. The government-sponsored funding agency Genome BC has invested more than Can\$273 million (US\$236 million) on biomarker research at academic institutions — ranging from identifying cancer-causing mutations to using genetic fingerprints to reduce adverse drug reactions in children — and on developing platforms combining bioinformatics, proteomics and genetic mapping.

Although scientific talent is abundant in the Pacific Northwest, there is a constant demand for those with clinical, regulatory or quality-control skills. Even more important for start-ups is to find a seasoned director to guide the business. "Finding skilled management remains the biggest challenge here," says Gabe Kalmar, executive director of operations at Genome BC.

The Pacific Northwest offers vibrant career options. Former University of Washington president Lee Huntsman's take on Seattle goes for the whole region: "It's big enough to be stable, but it's not at all set in its ways."

Virginia Gewin is a freelance science writer based in Portland, Oregon.

WEB LINKS

University of Washington
 ♦ www.uwashington.edu
 Fred Hutchinson Cancer Research Center
 ♦ www.fhcrc.org
 Seattle Biomedical Research Institute
 ♦ www.sbri.org
 Path
 ♦ www.path.org
 Genome BC
 ♦ www.genomebc.ca
 Leading Edge BC
 ♦ www.leadingedgebc.ca
 Oregon Health & Sciences University
 ♦ www.ohsu.edu
 ONAMI
 ♦ www.onami.us
 PNNL/UW Center for Nanotechnology
 ♦ www.nano.washington.edu

MOVERS

Neal Copeland and Nancy Jenkins, principal investigators, Institute of Molecular and Cell Biology, Singapore



1985-2006: Director, mouse cancer genetics programme, and head, molecular genetics, oncogenesis section, National Cancer Institute, Frederick, Maryland (Copeland).

Head, molecular genetics, development section (Jenkins)

1983-85: Associate professors, University of Cincinnati College of Medicine, Cincinnati, Ohio

Neal Copeland and Nancy Jenkins are a remarkable scientific double act, with some 700 shared publications in the decades since they met as postdocs at Harvard Medical School. They made their names working on mouse models of human disease at the Jackson Laboratory in Maine.

Now, after 20 years at the US National Cancer Institute (NCI), where they've built up a colony of 20,000 mice, this husband-and-wife team is moving to Singapore's Institute of Molecular and Cell Biology (IMCB). What's pulling them away from such a comfortable niche, in their 50s?

For a start, Singapore is investing heavily in bioscience, notably in the construction of Biopolis, a seven-building biomedical research complex. At the IMCB, which moved to new facilities in Biopolis last year, Copeland and Jenkins will be setting up a new cancer lab.

"We will be on the ground floor of something new and exciting," says Copeland. Decades of working together have created an easy-going harmony in which the two weave in and out of each other's sentences. Their closeness and many visits to Asia make the move less daunting.

"We've lived and travelled and worked together for more than 25 years," says Copeland. "We share an office..."

"Quite a small one..." says Jenkins.

"Most people would probably be divorced by now! You have to be compatible..."

"That doesn't mean we never yell at each other..."

"But it's worked out well..."

"... for us," Jenkins concludes.

They've made surprising moves before. They were the first molecular biologists to go to Jackson, where they fused molecular biology with mouse genetics. That led to a greater understanding of how cancers develop.

"The move there was risky," says Jenkins. Copeland adds: "A lot of our friends thought we were throwing our careers away: 'Why go into mouse genetics?'"

"They thought we'd disappear into the backwoods..."

"But it was the best thing we did. It shaped our careers."

Moving to the NCI was the next best thing, they say, and they're hoping Singapore will be the third.

Jenkins advises young scientists not to underestimate their fields. Developments such as transgenic mice and the mouse genome, which seemed like science fiction in 1980, greatly increased her and her husband's success.

"Have fun too," they add. "When it stops being fun, we'll probably retire."

Janet Wright

SCIENTISTS & SOCIETIES

Learning to teach

The role of teaching in the development of academic careers is often undervalued. Pre-eminence is afforded to research, and administration places growing demands on academic staff. Yet good teaching is part of a university's purpose, and fundamental to the fully rounded development of any academic practitioner.

At the University of Oxford, with funding from the Higher Education Funding Council for England, we are developing a Centre for Excellence in Preparing for Academic Practice. This will support postgraduate research students and contract research staff who intend to pursue academic careers and want to develop teaching skills.

Such support includes introductory teaching development courses and the opportunity to acquire a qualification. Students will have access to mentors and will participate in different modes of teaching. They will also be provided with the technological means to record and analyse their own pedagogical conceptions and techniques.

These programmes will be available to Oxford postgraduate students in all disciplines by 2010. Some departments, such as biochemistry, will bring their versions on-stream this year.

The centre's approach is grounded in the view that the development of teaching ability is itself a scholarly and

research-based activity, from which best practices can be informed and developed. During its five years of funding, the centre will aim to assist research postgraduates at Oxford, and to disseminate its research findings for use by wider audiences.

One critical research area concerns the ways in which strengthened teaching capabilities may enhance research skills and productivity. If it can be shown that, by professionalizing teaching as well as research competencies in the early stages of academics' careers, their abilities in both areas are improved, then there will be significant implications for young researchers, as well as for university managers seeking to promote research output within their institutions.

But even if the relationship between teaching and research turns out to be ambiguous or unproven, strong teaching standards will still be essential for the recruitment and retention of students. Sharper analysis of the teaching/research dialectic should prove valuable in shaping both career development and overall policy direction within universities.

Keith Trigwell is director, and Richard Arnold is manager, of the Centre for Excellence in Preparing for Academic Practice. For more information e-mail keith.trigwell@learning.ox.ac.uk

GRADUATE JOURNAL

PhD, take two

My first PhD attempt was not the experience I had imagined. Although I made great friends both in and outside the department, I was in the wrong lab. I had my own funding but felt like a slave: I could do only experiments I was told to do and could not choose my own course of research. I was forbidden to take a class in evolution because it wasn't "related enough". I founded a graduate students' association but was told to spend more time at the bench. Feeling demoralized and insignificant, I quit last year, six months after starting my PhD.

At first I was sent reeling: was this the end of my career? Was I a terrible scientist? Should I study something different? Genetic counselling, medicine, even philosophy? After a short stint as a waitress back home in Scotland, I went to Montreal to work as a research assistant in a small lab on the evolution of sex. I got to read as much as I wanted, design experiments and talk to my supervisor whenever I liked. No, I wasn't a bad scientist, and yes, science still excited me. I realized that I wanted to give a PhD another try.

I got married and returned to Britain, but the lessons I'd learnt were not forgotten. When deciding on a new lab, I made sure that both the department and supervisor were flexible, encouraging and accessible.

Here I am six months into my new PhD. This time last year I was desolate. Now I am full of hope. What will this year bring? Watch this space.

Mhairi Dupré is a first-year PhD student in evolutionary developmental biology at the University of Oxford, UK.

Don't imitate

Wait for the real thing.

Gilles Amon

The obvious one to choose, but the most difficult one to conceptualize. The last bottle of the evening, of course, has to be her. The diva. The *grande dame*.

It will seem unreal to get in touch with the Palmer '07 one more time — maybe for the last time. Distant memories of her playful youth will surface; and of the painfully long time it took her to mature, or rather, to transform. Only when she was well into her third decade did she start truly to intrigue us, at a point when we had almost stopped believing that there was something extraordinary in this wine. And extraordinary she is.

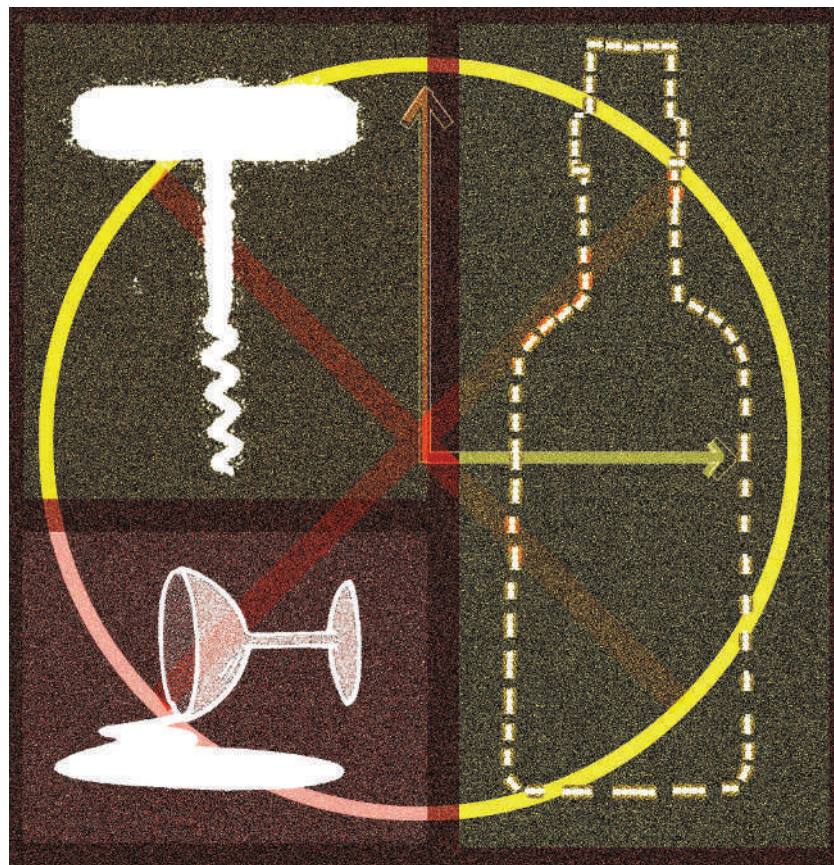
Of course, we could have paid more attention to certain details. A few did. But we opened the young bottles far too carelessly, too often and too soon. Had we only had the slightest idea of what a splendid future the 2007 was to experience, we wouldn't have put a premature end to the lives of so many bottles, depriving each one of the peaks to come.

But we did — like most of us so-called real wine aficionados. The most painful thing about having emptied so many bottles decades ago — rather than celebrating them now, when the wine is at the top of its game, and has left standing everything that was ever put in a bottle — the most painful thing is that, had we shown more foresight, it would not be my last bottle of beloved Palmer '07 that we will open tonight.

The biggest worry, of course, is — as with all bottles that are not glued — the cork. But I trust it will be fine. The bottle got a new cork in 2037, *au château*. How on Earth they managed to find natural corks 12 years after the European Union banned them, I have no idea. It seemed, back then, fairly early to be replacing the cork, but it was really the last opportunity to do so.

Not that much can be said against gluing the bottles (thankfully, there still are bottles around that need to be properly sealed). Far beyond the environmental aspect, the glue is just perfect nowadays. Many don't even own corkscrews any more. (I always held on to mine, like scientists who will never ditch their pencils.) The glue never, ever leaks. Never. Easily ripped off — and easily resealed, if you like that kind of thing.

After all, resealing is still better than pouring away your 2003 d'Yquem or 1945 Mouton, once you've had enough for the evening. This has become such a bad



JACEY

habit. But there are hundreds of other reasons to dislike imitations. They cannot be distinguished from the originals, of course. Who could forget that, 20 years ago, Gabriel Robertson — the elder — was unable to identify a single 'fake' in a series of 24 wines, which included some of his personal favourites? A shock.

Back then 'test-tube wines' were still produced under the Dior name, I think, before the branch was taken out of the group a few years later to form Vinor. No one can deny that it was the heavy involvement of the perfume industry that drove the early, purely scientific attempts at 'copying' famous vintages — for archival reasons, they said — away from their initial innocence, and towards a monstrous industry.

It wasn't long before everyone owned every vintage of Pétrus ever produced. A period of unspeakable snobbism began: "I don't ruin my burger with a 99-point wine." Give me a break. Even today, I find it devastating that everybody ran with the idea — and most still do. At least that cruelty of 'instant wines' never got off the ground. But nonetheless, countless wineries were faced with ruin, most of them

small, independent concerns; but also some big names. Fortunately, the majority of the Grands Crus survived; at least, those of the 2021 reclassification.

Oh yes, you can still find them, the pearls, but you have to sniff them out. And new stars have appeared, but only a handful, produced by enthusiasts who, audaciously, set out to celebrate wines in their entirety with the same intensity as us. Lucky us, who, unperturbed, kept on searching for real wines, 'vinified wines', as they call them these days.

While others stick their noses in 'Best of Wine' collections from the corner shop, we are the ones who can see great vintages develop, as they become, well, ready. The 2029s, the 2039s and, above all, the 2007s. Only we witnessed history unfolding, snapshot by snapshot. Often you don't have much of an idea what will have happened since the last time you looked. Some of the best stories are still being written. And much is still to come, many surprises, I dare say.

Well, enough talking. Let us celebrate the diva.

Gilles Amon lives in London and enjoys wines. ■